DOCUMENT RESUME

ED 117 175 95 TM 005 031

AUTHOR	Quilling, Mary Rintoul
TITLE	A Data Analysis Approach to Evaluating Achievement Outcomes of Instruction. Technical Report No. 338. Report from the Project on Conditions of School Learning and Instructional Strategies.
INSTITUTION	Wisconsin Univ., Madison. Research and Development Center for Cognitive Learning.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	TR-338
PUB DATE	Jun 75
CONTRACT	NE-C-00-3-0065
NOTE	153p.
EDRS PRICE	MF-$0.76 HC-$8.24 Plus Postage
DESCRIPTORS	*Academic Achievement; Comparative Analysis; *Curriculum Evaluation; *Data Analysis; Elementary Secondary Education; Mathematical Models; *Predictor Variables; *Research Methodology; Statistical Analysis

ABSTRACT

The purpose of the present study is to demonstrate the utility of data analysis methodology in evaluative research relating pupil and curriculum variables to pupil achievement. Regression models which account for achievement will result from the application of the methodology to two evaluative problems--one of curriculum comparison and another exploring the relationships between achievement and instructional processes in different schools implementing the same curriculum. Evaluative studies focusing on such questions should yield more information about pupil achievement than evaluations following other models when the following practices are reflected in the design and execution of the study: (1) several dimensions of the curriculum, including material and instructional process aspects, are represented in the set of independent variables; (2) curricular and pupil variables are chosen whenever possible from those conceptualized by educational researchers and known to have a likely relationship to achievement; (3) direct measurements on all variables are incorporated in the data set rather than categorical representations of variables, as in a factorial design; (4) the shape as well as the location of the distributions of pupil achievement before and after instruction is represented in the analyses; (5) the techniques of the data analyst guide the model development process. These recommendations result from a review of substantive and methodological literature. (Author/BJG)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED117175

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Technical Report No. 338

A DATA ANALYSIS APPROACH TO EVALUATING ACHIEVEMENT OUTCOMES OF INSTRUCTION

by

Mary Rintoul Quilling

Report from the Project on
Conditions of School Learning and Instructional Strategies

Wisconsin Research and Development Center
for Cognitive Learning
The University of Wisconsin
Madison, Wisconsin
June 1975

TM 005 031

Published by the Wisconsin Research and Development Center for Cognitive Learning, supported in part as a research and development center by funds from the National Institute of Education, Department of Health, Education, and Welfare. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that agency should be inferred.

Center Contract No. NE-C-00-3-0065

WISCONSIN RESEARCH AND DEVELOPMENT CENTER FOR COGNITIVE LEARNING

MISSION

The mission of the Wisconsin Research and Development Center for Cognitive Learning is to help learners develop as rapidly and effectively as possible their potential as human beings and as contributing members of society. The R&D Center is striving to fulfill this goal by

- conducting research to discover more about how children learn
- developing improved instructional strategies, processes and materials for school administrators, teachers, and children, and
- offering assistance to educators and citizens which will help transfer the outcomes of research and development into practice

PROGRAM

The activities of the Wisconsin R&D Center are organized around one unifying theme, Individually Guided Education.

FUNDING

The Wisconsin R&D Center is supported with funds from the National Institute of Education; the Bureau of Education for the Handicapped, U.S. Office of Education; and the University of Wisconsin.

ACKNOWLEDGMENTS

Two professors have been instrumental in shaping the direction of my professional development and in supporting the research represented in this thesis. Professor Frank B. Baker, who served as major professor throughout my graduate program, encouraged and guided the preparation of the thesis. I am especially appreciative of his helpful criticism and of his responsiveness to my interests in applying methodological techniques to current educational issues. The commitment of Professor Herbert J. Klausmeier to improving the quality of education and my opportunity to pursue that goal working under his direction at the Wisconsin Research and Development Center for Cognitive Learning influenced the choice of the thesis topic, but more generally led to my concern for educational applications of quantitative techniques. For the numerous professional experiences he has accorded me and for the privilege of working with him, I am indeed grateful. My thanks are also extended to Professor Lawrence Hubert, Professor M. Vere DeVault, and Professor John Gurland for serving as members of the committee.

The administrative and technical staff of the Wisconsin Research and Development Center for Cognitive Learning generously supported the conduct of this study. I am particularly indebted to Deborah Stewart, Edward Haertel, Anne Buchanan, Pamela Klopp and Patrick Lane for their assistance in various aspects of conducting and reporting the study and for their numerous personal kindnesses in ensuring a timely completion of the research. Appreciation is extended also to the other Center staff members, too numerous to mention, who contributed of their talents. Finally, the forbearance of the staff of the Environmental Education Project, as leadership and secretarial services were diverted from its mission, deserves recognition. To my colleagues and friends at the Center, thank you.

TABLE OF CONTENTS

Chapter Page

Acknowledgements iv

List of Tables ix

List of Figures xi

I. The Nature of the Problem 1

  Review of the Literature 7
    Instructional Models 7
    Evaluative Methodology 15
    Data Analysis Applied in the Regression Context . 20
      Analysis of Residuals 22
      Outliers 23
  Statement of the Problem 24
  Summary 26

II. Regression and Data Analysis 29

  The Regression Model and Underlying Assumptions . . 30
  Goodness of the Data in Other Regards 33
  Evaluating the Regression Equation 37
    Assessing Overall Adequacy of the Model 37
    Analysis of Residuals 40
      Distribution of Residuals 40
      Residual Plot Against $\hat{Y}_i$ 41
      Residual Plot Against Terms in X 43
      Residual Plot Against Factors External to the Model 43
      Outliers 43
  Application of Procedures in the Present Study . . . 45

III. Variables in the Regression Model 51

  Dependent Variables 53
  Independent Variables 65
    Pupil Variables 66
    Curricular Variables 67

Chapter Page

Curriculum Decision Unit Size 70
Rate Adaptiveness 72
Structural Granularity of the Content 74
Content Diffficulty 74
Interactions 76
The Unit of Analysis 76
Statistical Descriptors of Classroom and Grade Level Groups 78
Data Sets Entered into the Statistical Analyses . . 81

IV. Development of the Predictive Equations 85

Feasibility Study 86
Preliminary Regression Analyses 86
Statistical Goals for the Study 87
Model Development in the Curriculum Tuning Evaluation 90
Inspection of the Data 91
Fitting an Equation to Predict the Mean 93
First Regression Analysis 93
Regression Analysis with Modified Data 98
Regression Analysis with New Variables 100
Model Verification 104
Fitting an Equation to Predict the Standard Deviation 105
Fitting an Equation to Predict the Skew 108
First Regression Analysis 108
Regression Analyses with New Variables 109
The Set of Predictive Equations for Reading Achievement 113
Model Development in the Curriculum Comparison Evaluation 114
Inspection of the Data 114
Fitting an Equation to Predict the Mean 118
Fitting an Equation to Predict the Standard Deviation 121
Fitting an Equation to Predict the Skew 123
The Set of Predictive Equations for Mathematics Achievement 125

V. Conclusions and Discussion 127

An Appraisal of the Effectiveness of the Techniques for Purposes of Evaluation 128
Curriculum Tuning Study 128
Curriculum Comparison Study 131

Chapter Page

Appraisal of the Findings in Light of the Goals of the Study 133
The Utility for Evaluation Purposes of the Variables in the Study 133
Data Analysis as a Statistical Technique Useful in Evaluation 134
A Prototype for Evaluative Research on Educational Achievement 136

References 139

LIST OF TABLES

Table		Page
1	Analysis of variance to check lack of fit	39
2	Criterion measures of pupil achievement used in the "Curriculum Tuning Study" of the *Wisconsin Design for Reading Skill Development* (WDRSD)	63
3	Predictor measures of pupil achievement used in the "Curriculum Tuning Study" of the *Wisconsin Design for Reading Skill Development* (WDRSD)	68
4	Basic sets of criterion and predictor variables . . .	82
5	Summary of exploratory analyses	88
6	Descriptive statistics for basic variables in Model 1	91
7	Correlation matrix for basic variables	92
8	Arrangement of observations in the X space	94
9	Analysis of variance--first regression	95
10	Residual analysis for first regression	96
11	Analysis of variance--modified data	98
12	Coefficients and related statistics for regression with new variables in the data set	101
13	Analysis of variance with additional variables in the data set	101
14	Summary of regression analyses to predict D_7:m	103
15	Comparison of predictive efficiency of model using original data and a data set with a different predictor measure of achievement	104

Table Page

16 Coefficients and related statistics for the first regression to predict the standard deviation 106

17 Analysis of variance--first regression on standard deviation 107

18 Comparison of original regression equation and one adding a crossproduct term 111

19 Comparison of regression equations with one and two interaction terms added 112

20 Analysis of variance--final regression to predicting skew 112

21 Regression equations predicitng the distribution of reading achievement 114

22 Descriptive statistics for basic scaled variables in the curriculum comparison study 115

23 Correlation matrix for basic variables in the curriculum comparison study 117

24 Analysis of variance--first regression to predict mean arithmetic performance 119

25 Data for two equations which predict the class mean in arithmetic achievment 121

26 Descriptive information regarding equations successively developed to predict class standard deviations in arithmetic 122

27 Coefficients and related statistics for the final equation to predict skew 123

LIST OF FIGURES

Figure		Page
1	Possible patterns in the distribution of residuals against the value of Y fitted through a regression equation	42
2	Sequence of data analysis procedures	47
3	Schema of a hypothetical relationship between the content of two programs and a criterion test	61
4	Plot of residuals against $\hat{Y}_{m_1}$ with X_5 in regression equation	97
5	Plot of residuals against X_1 with X_3 and X_5 in predictive equation	99
6	Plot of residuals against three school types with X_3 and X_5 in predictive equation	100
7	Plot of standardized residuals for regression equation with X_5, X_3, and X_7 entered	102
8	Plot of residuals against X_1 with X_3, X_5, and X_7 in the equation	103
9	Plot of residuals against X_4 for first equation for Y_s	107
10	Plot of residuals from first equation predicting skew against X_5, rate adaptiveness	109
11	Plot of residuals against the crossproduct of X_3 and X_5 for the skew predictive equation	110
12	Plot of residuals from the first equation predicting skew against the X_4X_5 crossproduct	110
13	Plot of X space for curricular variables in the mathematics study	118

Figure Page

14 Plot of residuals for 19 classes after terms in X_1, X_2, X_6 and X_2X_6 enter the equation to predict mean class achievement 120

15 Plot of residuals from initial equation to predict skew in the distribution of arithmetic scores, with one variable entered in the regression equation 124

Chapter I

THE NATURE OF THE PROBLEM

In recent years there has been a proliferation of new curricula developed and published. Curricula are being produced by federally funded laboratories and by research and development centers in university settings and other institutions, as well as by local school districts. Both developers and consumers of the new products are faced with the need to evaluate the effectiveness of the curricula, particularly with respect to pupil achievement. In the case of a developer, data from a summative field test usually provide limited evidence on product effectiveness. The results, for instance, frequently do not demonstrate the comparative effectiveness of the target alternative curricula, nor do they generalize to all the pupil subpopulations of interest. Nevertheless, a question of major interest to the professional staff faced with a curricular adoption decision is how one product compares with another when used with student groups similar to those in the district's schools. Since the limited information that is usually available to the adopter is insufficient for him to make a reasoned choice for his schools, many school districts choose to appraise pupil performance and other outcomes after a product has been installed and utilized for a year or two. In this case, the achievement of pupils assigned to the curriculum is often compared with the achievement of pupils assigned to one or more other curricula in use previously or concurrently. Too often expectations for the new curriculum are not fulfilled, or mixed results undermine a district's commitment to make the curriculum work. Most curriculum comparisons, in short, have not facilitated decisions that resulted in significant educational improvement (Astin & Panos, 1971).

Instead of merely comparing curricula, evaluation might well focus on discovering the conditions associated with various outcomes, as Cronbach (1963) suggested a decade ago, so that characteristics of implementation may be modified or instruction "tuned" to maximize learning. This approach to evaluation is suitable either for an institutional developer undertaking summative evaluation of a near-completed product or for a school district appraising the effectiveness of its implementation of the program. An important feature of the approach is the potential for constructively using the information yielded to improve the present or future curricula. Like most other evaluations, one focusing on the relationship between outcomes and conditions must usually be conducted under financial and other constraints. Limitations are placed on the number of schools that can participate in the field test and on the duration of the field test. The amount and kind of data collected are constrained by resources, school policy, and teacher attitudes. Routine analysis of data, coupled with these constraints, leads to what is often an unproductive outcome. Instead, measurement of the conditions of instruction and an analysis of the data that is tailored to the nature and amount of the data may yield more useful information.

The purpose of the present study is to illustrate how data analytic techniques may be usefully applied to curriculum evaluation carried out within the framework suggested above. The particular application involves developing regression models that account for school achievement in terms of curricular and pupil variables and their interactions. These independent and dependent variables are similar to the inputs, operations, and outputs of the Astin and Panos program evaluation model (1971) that has been applied to the evaluation of college instruction (Astin & Panos, 1969) and the variables of Cooley's (1971) correlational model--student entering behaviors, dimensions of instructional treatments, and end-of-year student achievement--that has been utilized with field test data from Individually Prescribed Instruction curricula. While the overall approach to evaluation is thus not new, certain of the variables used in the present study and the data analysis procedures have not been applied seriously in the context of evaluative research.

The stimulus for the study comes from several areas of both substantive and methodological research. Educational researchers, curriculum theorists, evaluators, and statisticians have constructs useful in studying the relationship of school achievement to curricular variables, pupil characteristics, and their interaction. An integration of the constructs in an evaluative study should enrich the findings. A brief overview of some of the substantive and methodological considerations will suggest more specifically the variables of interest, the usefulness of a regression approach, and the relevance of data analytic techniques.

In the substantive domain, Bruner's (1964) call for a theory of instruction has stimulated both consideration of variables which differentiate such a theory from learning theories and development of some instructional models. Recently Lohnes (1972a) and Atkinson (1972) have expressed interest in a theory with marked quantitative characteristics and have utilized mathematical models in developing such a theory. Using a regression approach, Suppes (1967) and others have presented models of school learning, but such an approach has only recently been applied to curriculum evaluation *per se* (Cooley, 1971).

In another substantive vein, the knowledge collected during this century on individual differences (e.g. Cook, 1951) has been applied to instructional development. The proposition that pupil achievement will improve when individual differences are taken into account underlies the current commitment of developers and practitioners to individualized instruction. Field tests of individualized curricula, however, have generally failed to yield dramatic results (Gage & Unruh, 1967) and the ultimate promise of individualized instruction may be fulfilled only when more progress has been made in the study of aptitude-treatment interactions, suggesting how instruction should be differentiated for pupils of various traits (Berliner & Cahen, 1973). This field of learning research may provide information useful in the development of a theory of instruction, if variables can be identified that both interact significantly with aptitudes and can be manipulated in the classroom.

With respect to methodology, the major evaluation prototypes of Tyler (1942), Stake (1967, 1972, 1973; Stake & Gjerde, 1971), Stuffelbeam (1969), and Scriven (1967, 1972) and special-purpose models such as that offered by Taba (1966) do not integrate the concerns of instructional theorists and learning modelers, nor are the evaluation models easily adapted to the particular propositions of individualized instruction and aptitude-treatment interaction research. Nonetheless, some of the key ideas in these formulations--including Tyler's focus on specified instructional objectives, Scriven's analysis of the function and form of summative evaluation, Stake's concern for description, Taba's interest in instructional variables that affect achievement, and Stuffelbeam's goal of facilitating decision making--are related to the present study. To accommodate these ideas in studying the relationship between achievement and independent variables, the data sets of the evaluator need to be expanded in terms of the number and kind of variables represented in the evaluative design.

The measurement of such variables and the statistical design of a study are related methodological problems. Typically in studies of school achievement, curricula are conceived of holistically; two or more curricular treatments are represented nominally in a statistical model that accounts for pupil achievement. Curricula, however, differ on a number of dimensions related to content and instructional characteristics, and many of these dimensions can be measured on an interval scale. The classroom setting of curriculum evaluation virtually ensures differences in instruction from classroom to classroom; in this case nominal representation of the instructional treatment as a single entity masks potentially useful information regarding the effect of such differences on achievement. Furthermore, Cronbach and Snow (1969) have pointed out that a research design using indicator variables or factors, instead of quantitative measurements, is not maximally sensitive to the interactions of interest to aptitude-treatment researchers. Thus, it is desireable to represent curricula in a statistical model in terms of quantified attributes. In educational research the statistical model is ordinarily linear and, for a data set whose independent variables are represented by actual measurements, regression analysis is the appropriate methodology. The more commonly used factorial designs with corresponding analyses of variance are a special case of the linear model and regression analysis. In either case, however, the data gathered in a school setting are "messy" and special tools of analysis should augment the conventional analyses indicated by the statistical model.

The application of statistical techniques to "messy data" was a subject in Tukey's (1962) treatise on data analysis. His suggestions stimulated the development and use of techniques, in addition to those he discussed, to aid the applied researcher in handling empirical data. These techniques were designed to deal with problems such as "lurking" variables, outliers, the satisfaction of statistical assumptions, and evaluation of the adequacy of a statistical model. Despite their applicability to problems in education and psychology, the techniques are rarely used in these fields. Rather, the experimental designs and statistical analyses employed continue to follow the Fisherian tradition of stand-alone experiments with factorial designs. The complex, multifaceted nature of educational problems, including those posed for a curriculum evaluation, requires new techniques that allow for iteration, sensitivity to the characteristics of the object being studied and their representation in fallible measurements, introduction of new variables into the statistical model and rejection of insignificant variables. When properly exploited, data analysis should yield insights into the problems to which it is applied in addition to results from a conventional analysis. How the flexible, open-ended approach of the data analyst might contribute to the study of instructional processes and outcomes will be illustrated in the thesis. The application of data analysis to curriculum evaluation is worthy of exploration because the techniques promise to be a powerful addition to the evaluator's methodological repertoire.

In the remainder of this chapter the substantive and methodological themes mentioned above will be expanded upon.

REVIEW OF THE LITERATURE

Instructional Models

The likelihood that evaluative studies using data analysis methodology will yield fruitful results is greater when the instructional situation is carefully conceptualized. In particular, the variables represented in the model should be selected in light of research findings or other strong justification. The data analysis approach, however, permits modification of the initial model on the basis of observations or information gathered during a study. Thus, a data analysis approach promises to contribute toward model development and refinement. Models, which are systems for representing empirical events and relationships, contribute to the construction and application of theory (Lachman, 1960). Model development is therefore an appropriate step toward Bruner's (1964) goal of a prescriptive theory of instruction that "sets forth rules concerning the most effective way of achieving knowledge or skills. . . . and providing a yardstick for criticizing or evaluating any particular way of teaching or learning [p. 306]." His call has stimulated a wide array of approaches by applied and basic researchers that range from descriptive, unvalidated accounts of the learning process to models specifying rules for presentation of content. Carroll (1963) perhaps expressed best what characteristics such models should have:

> What is needed is a schematic design or conceptual model of factors affecting success in school learning and of the way they interact. Such a model should use a very small number of simplifying concepts, conceptually independent of one another and referring to phenomena at the same level of discourse [p. 723].

The models of Suppes (1967) and Atkinson (1972) are illustrative of substantively and methodologically sound approaches which may lead to an instructional theory with marked quantitative characteristics. Suppes, using a regression approach, has developed mathematical models of arithmetic and geometry learning that account for group achievement in light of characteristics of the mathematical problems. Atkinson suggested that a theory of instruction should have a dynamic, interactive quality in which an instructional model is utilized to make instructional programming decisions to optimize the learning of an individual pupil.

Lohnes (1972a) similarly advocated the quantitative approach and focused his efforts on selecting instructional variables on the basis of known quantitative relationships between such independent variables and results, as well as on measuring instructional characteristics more accurately. DeVault and his colleagues (1973), in order to analyze individualized programs, have developed descriptors for both curricular materials and instructional processes measured at a classroom level. His approach to rating the extent to which a particular variable is represented in a program is consonant with Lohnes' (1972a) observation that "treatments are more realistically viewed as differing in degrees of emphases, rather than in absolute kind [p. 1]." Additional explicit references to identifying instructional parameters that should be represented in a theory of instruction exist in the literature on instructional models. Interest has been shown primarily in characteristics inherent to a subject matter (Suppes, 1967) and in instructional process variables (Atkinson, 1972; Carroll, 1963; Cooley, 1971; Lohnes, 1972a).

Little attention has been given by evaluators to characteristics of curriculum materials, particularly their organization. Despite the fact that single facets of curriculum materials are manipulated at several levels for experimental purposes, seldom is the degree to which particular curriculum materials possess a given characteristic considered in school adoption or assignment decisions. The holistic view of a curriculum and the corresponding nominal treatment of a program in evaluation designs, however, may result from a lack of explicit information on the degree to which a trait exists in given materials rather than an actual absence of identifiable, even measurable, characteristics. Walbesser (1972), for instance, has proposed a measure of "puissance" for the kinds of learning objectives in comparable mathematics programs, and aptitude-treatment interaction researchers have devised a variety of measures of task or content structure (Berliner & Cahen, 1973). Few variables measuring the organization of content and other material dimensions, however, are in use by applied researchers or evaluators working in the school situation. Nevertheless, the fact that a curriculum has organizational and format aspects, as well as related instructional strategies, suggests that models predicting achievement should include variables relevant to each source of variation. The curriculum materials involved in comparative studies, for instance, might well differ on dimensions of the kind discussed.

In addition to considering the variables that might be represented in quantitative instructional models, further attention to the technical form of an instructional model is warranted. From the viewpoint of the applied mathematician or statistician, there are three kinds of models: functional, control, and predictive (Draper & Smith, 1966). The models differ in the degree to which the equations fit the data and therefore in the purposes that each model serves. The functional model is characterized by often complex, higher order equations that give a precise degree of fit to the data, permitting a good understanding of what effects the response. Psychological and educational phenomena are not presently well enough understood to be easily represented by true functional models. Such a model is a goal toward which the social scientist strives as he develops a theory that is useful in understanding, controlling, and predicting the response. Bruner's conception of a theory of instruction might ultimately be represented best by a functional model, but first steps toward the theory utilize less sophisticated models.

Models which focus solely on variables under the control of the user are needed to optimize or otherwise manipulate a response. Developing a control model requires experimental manipulation of each variable at levels over their possible range, so that the effects are known singly and jointly. Control models appear to be useful in industrial situations and in fields such as physical science. In education and psychology, control models have been utilized for clearly conceptualized and narrowly defined processes such as the presentation of programmed instructional frames (Smallwood, 1962) to maximize individual learning.

To date, however, psychological and educational phenomena are most tractable to predictive models. The characteristics and utility of such models have been aptly described by Draper and Smith (1966):

> When the functional model is very complex and when the ability to obtain independent estimates of the effects of control variables is limited, one can often obtain a linear predictive model which, though it may be in some senses unrealistic, at least reproduces the main feature of the behavior or the response under study. These predictive models are very useful and under certain conditions can lead to real insight into the process or problem. . . . The predictive model is not necessarily functional and need not be useful for control purposes. . . . If nothing else, it can and does provide guidelines for further experimentation, it pinpoints important variables, and it is a very useful variable-screening devise [p. 235].

Multiple regression techniques are appropriately utilized in developing predictive models, and the existence of "messy data" does not prevent a successful conclusion to the development of the model, given the techniques of data analysis. While the predictive model is the lowest level of the three models, developing such a model to relate achievement to curricular and pupil variables is a suitable beginning point.

Most proposed or extant models of school learning are, indeed, predictive models. Carroll's (1963) relatively early model was designed to predict school achievement in terms of five temporal variables, three of which pertained to individual characteristics and two to instructional circumstances. Degree of learning was considered a function of the total time spent in learning in relation to the time needed. The obvious motivation for a model relying solely upon the time indices is that all variables are measured on an interval scale. Besel (1971) applied Carroll's predictive model to the allocation of CAI time given the classroom time and equipment contraints, and thereby "upgraded" it to a control model. Although Carroll's model has an apparent conceptual tie to individualized curricula that pace instruction differentially, its application to predicting achievement is unwieldy because of the abstraction of several of the variables and difficulty in gathering certain of the measures. Some of the variables are not directly measurable and seem more appropriate for use in a finely tuned research study than in curriculum evaluation. Nonetheless, scant attention was given the "time provided" variables in the educational literature before Carroll's model appeared, and these are significant variables in individualized programs now in use. Also, Carroll's model has a high degree of generality to learning in various subject matters.

Atkinson (1972) provides an illustration of a simple predictive model that is used to guide item-presentation decisions in a CAI system. Taking into account the individual's and group's learning histories, Atkinson predicted the probability of success on alternative items so that an item having the highest probability of being learned next might be selected for presentation. Again, the approach generalizes to different subject matters, though not to as wide a variety of instructional modes as does Carroll's.

A model applied to a highly specific task with variables intrinsic to the subject matter is provided by Suppes (1967). With the intent "to combine the logical structure of the subject and the actual performance of a student [p. 6]" he used a linear model to predict with good fit the error rate on a given arithmetic computation item in terms of its number of computational steps and the magnitude of the numerals. Similarly, his model for geometry performance was based on properties of the geometric figure, such as right angles and number of sides. The obvious shortcoming of this approach is that a new model has to be conceptualized and developed for each content area.

Models that are functionally similar to those described above have not yet been developed by researchers investigating aptitude-treatment interactions (ATI). ATI work is presently focused on screening aptitudes and other traits that interact disordinally with treatment dimensions. The discovery of significant variables that might ultimately appear in a parsimonious model of school learning has been hampered by methodological practices, such as the use of gain scores as dependent measures, the predominant application of inferential rather than descriptive statistical procedures, and research designs calling for the blocking of independent variables (Cronbach & Snow, 1969). The usual ANOVA, in the latter case, has a larger error component attributable to within cell differences, than does the more efficient regression approach using observed scores. Furthermore, application of the conservative Bracht-Glass (1968) strategy for detecting disordinal interactions has likely constrained the identification of some that would have been recognized with a more liberal decision rule.

Despite the meager findings to date (Cronbach & Snow, 1969; Bracht, 1970; Berliner & Cahen, 1973) and almost total absence of ATI research in an evaluative context, the ATI approach is relevant for this study. First, although many of the ATI findings are inconclusive at the present, ATI research is a fruitful source of curricular variables already well defined and potentially significant. Second, the methodology of ATI research proposed by Cronbach and Snow is analogous in many respects to that of the present study. Not to be overlooked is the possibility of a serendipitous finding in the evaluation of curricula used over periods longer than those during which ATI researchers typically manipulate the variables of interest.

The preceding discussion of models and their application to school learning has served to illustrate both the kinds of variables that are typically entered into equations and the kinds of problems that have been approached through modeling. The researchers and curriculum developers mentioned have shown a concern for finding important measurable variables that correlate with achievement and should therefore be included in models of school learning. Carroll's proposed model is the closest in its intent to that of the present study. Since the model was published, new emphasis has been placed on measuring a variety of characteristics of curricular and instructional-process variables. Data analysis offers techniques for screening these variables to determine whether they should be included in models to evaluate curricula in terms of achievement criteria.

Evaluative Methodology

Widely used evaluative prototypes, loosely defined as models, include those forwarded by Tyler (1942) and Stuffelbeam (1969) as well as a pervasively used general achievement testing strategy that gained favor in the 1920's. Tyler's formulation, which currently underpins the National Assessment of Educational Progress and the evaluation of countless objective-based curricula, calls for specification of instructional objectives, development of criterion-referenced tests for the objectives, and measurement of student progress toward the objectives. Tyler's model differs from the loosely defined testing strategy followed by many school districts in requiring a clearer definition of the content to be learned and in using criterion-referenced rather than norm-referenced tests. The relative merits of norm-referenced and criterion-referenced achievement tests are discussed in Chapter III. Since either may serve as a dependent or pupil-input variable in a quantitative model of instruction, the choice ultimately is a matter of deciding what kind of cognitive behaviors are of interest.

Stuffelbeam's model is addressed primarily to administrators and involves application of a systems approach in identifying alternatives, studying implications, and making decisions. It stresses rational decision making, taking into account a number of variables in the educational setting, and does not focus particularly on pupil variables. For this reason, the model has few implications for the present study.

Models due to Stake (1967, 1972, 1973; Stake & Gjerde, 1971) and Taba (1966) and Scriven's (1972) new model are less widely used, perhaps because they are more complex for educators to use or understand. Stake and Scriven are both interested in a wide variety of educational processes and outcomes. Stake's countenance and advocacy models as well as his newer construct of "responsive evaluation," focus on describing these variables and making value judgments about them. Scriven's goal-free model calls for evaluating all the ramifications of a curriculum instead of focusing just on that subset of objectives specified by the curriculum. Both of these evaluation theorists emphasize the multivariate nature of education. To each, appraisal of pupil achievement is but one aspect of evaluation. Taba does not take as comprehensive a view, but suggests that curricular variables be identified and systematically varied to discover relationships with outcomes and to find rules for developing new curricula. The significance of her concept of evaluation for the present study lies in its focusing on variables under the control of the developer or teacher.

Curriculum evaluation prototypes only indirectly suggest how the data gathered in applying each model should be processed. For some models, notably that of Stake, observations or verbal descriptions are gathered and rated by a few people. Scores, either from norm-referenced or criterion-referenced tests, may simply be summarized descriptively. Only for the Taba prototype is a statistical model necessarily used in the evaluation. For this reason, it is apparent that, although most evaluation prototypes provide guidelines for selecting, summarizing, and interpreting data, they do not apply the methodological procedures used in model development or refinement.

Contrasting in substance and methodology with the prototypes described are the writings on evaluation of Astin and Panos (1971), Cooley (1971), Cronbach (1963), and Lohnes (1972a). They conceive of evaluation as what Cooley calls "evaluative research," differing from other educational research primarily in its school setting and in the kinds of questions that are investigated. Cronbach (1963), in an article significantly entitled "Course Improvement Through Evaluation," expresses the group's common interest in understanding what variables affect performance and in obtaining information of value beyond the particular evaluation:

> Insofar as possible, evaluation should be used to understand how the course produces its effects and what parameters influence its effectiveness. It is important to learn, for example, that the outcome of programmed instruction depends very much upon the attitude of the teacher; indeed, this may be more important than to learn that on the average such instruction produces slightly better or worse results than conventional instruction. Hopefully, evaluation studies will go beyond reporting on this or that course and help us to understand educational learning. Such insight will, in the end, contribute to the development of all courses rather than just the course under test [p. 675].

Noteworthy is the difference between improving instruction in general through significant findings, as Cronbach suggests, and refining a task or technique as a result of findings in a formative evaluation (Scriven, 1967). The latter less formal evaluation is often conducted earlier in the development sequence than the evaluation foreseen by Cronbach and deals more with questions about small segments of a product. In contrast, the parameters that are investigated in evaluative research permeate almost all instructional situations. Lohnes (1972a, 1972b) has focused his attention on describing pupil characteristics and selecting a parsimonious number of traits for inclusion in an instructional model. He considered, for instance, the relative merits of intelligence scores and subject-matter pretest scores as pupil-imput measures and concluded that the former account for more variance in performance after instruction than do the latter. Cronbach and Snow (1969) similarly endorsed "the substantial predictive value of general mental tests in instructional search [p. vii]." Cooley (1971) proposed seven classroom conditions that are particularly relevant for Individually Prescribed Instruction (IPI): testing procedures, prescriptive (or individualized) practices, teacher skills (e.g. use of reinforcement), instructional materials actually used, allocation of time, space and its utilization and teacher's knowledge of the curriculum and of children. Lohnes (1972a) expanded on this list by adding some content and material variables. The singling out of traits and curricular variables as parameters to be related to achievement in evaluative research obviously presents the opportunity additionally to look for aptitude-treatment interactions.

A research orientation to evaluation entails consideration of appropriate methodology. Many facets of correlational methodology, including regression, canonical correlations, and factor analysis have been applied by Cooley and Lohnes in their evaluations. Using Wiley's (1970) suggestion, the unit-of-analysis problems in school research have been resolved more satisfactorily than in the past by using moments or cumulants of the distribution of scores from a classroom. Despite the availability of these more sophisticated statistical procedures, only the exceptional evaluator is going to be, by training or circumstance, in a position to use the techniques. Ultimately, the success of the evaluative research approach depends upon parsimony in its constructs and simplicity in its techniques. Data analysis as a methodology promises to meet these criteria, thus adding a more analytic approach to a set of prototypes from which the evaluator may choose.

Data Analysis Applied in the Regression Context

The applicability of regression analysis to evaluative research, modeling of school learning, and study of aptitude-treatment interactions has been established in previous sections. Regression analysis, furthermore, meets a practical requirement that the methodology for an alternative evaluation prototype be within the statistical repertoire of the evaluator.

Regression analysis in a variable-screening situation requires a strategy for determining which variables to include and exclude. Standard textbooks discuss the merits of well known strategies including stepwise regression, stagewise regression, forward selection, backward elimination, combinations of the preceding, and all possible regressions. Stepwise regression is favored by Draper and Smith (1966); the "all possible regressions" method by Daniel and Wood (1971), who propose an index for choosing among different equations. It is important to note that the use of different strategies yields different results.

The fact that evaluative researchers are in the stage of exploring new variables and building new models suggests that standard regression analysis be augmented by data analytic techniques. Data analysis is a statistical philosophy and methodology initially advanced and described by Tukey (1962) and not necessarily limited to the regression framework. He portrayed data analysis as a flexible, responsive approach to real world problems, utilizing the contributions of mathematical statistics to guide one's judgment rather than as an authoritative set of rules. According to his meaning, data analysis includes "procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data [p. 2]." In advocating relaxation of rules when judgment suggests, he states that "data analysis must be willing to err moderately often in order that inadequate evidence shall more often suggest the right answer."

Box,[1] Draper and Smith (1966), and Daniel and Wood (1971), in applying the philosophy of data analysis specifically in regression situations, stress the usefulness of iterative analysis and repeated experimentation or model verification steps. At each step of the iteration, the analysis suggested by the design of the study is initially carried out routinely and the assumptions are checked, new information, even hunches, represented in the statistical model for a reanalysis, and tentative conclusions reached which a subsequent experiment attempts to replicate. An array of statistical and graphic techniques, intuition, judgment, and knowledge of the phenomenon guide the analyst's interaction with the data. Initially the location of data points in the X space is inspected to determine whether the data are distributed over the range of interest. Consideration of the distribution of residuals and plots of residuals against the dependent variables follows the initial analysis and guides decisions to modify or reject outliers or to use transformations or higher order models in reanalyses. The essential feature throughout the iterative sequence leading to model refinement is the creative use and interpretation of data to guide the next step. Two of the more widely known and frequently used techniques are described below.

---

[1] Box, G.E.P. Course notes, mimeographed, 1966

Analysis of Residuals

A residual is the difference between an observed response, Y_i, and a response, $\hat{Y}$, predicted from each observation, X_i, by substituting in the regression equation least squares estimates of parameters. A simple regression model with one independent variable, X, and one dependent variable, Y, may be represented by the equation $Y = \alpha + \beta X$. Solution of the equation gives estimates a and b for α and β, respectively. Substitution of the estimates a and b gives for each observation X_i, the predicted value $\hat{Y}_i$. The fit between the model and the data may be examined by looking at each residual, $Y_i - \hat{Y}_i$. Individual residuals may be graphed or inspected individually, and various statistical tests can be applied to the residuals as a set.

Anscombe and Tukey (1963) described the techniques for examining residuals and making interpretations. Their techniques reveal problems of fit due to curvilinear relationships, a correlation between the fitted value of the dependent variable and the size of the residual, and the need for an additional variable, such as time, in the equation. Both half-normal plots and normal plots, which linearize normally distributed data when it is plotted on special grids, are useful in checking the assumption of normality and in detecting outliers. These and other techniques of analyzing residuals are a key feature of regression in the data analysis tradition.

Outliers

Extreme data points in small samples can have an undue influence on the estimates of the parameters in a regression equation, and consideration of whether such data points are "bad" values is encouraged. Two analytic techniques for dealing with such values are commonly used--trimming and Winsorizing. In the simplest application, trimming is simply a matter of discarding such bad values and proceeding with the data left. Winsorizing, named after Charles Winsor, involves changing the value of an outlier to the value of the nearest observation not considered suspect. Mathematical statisticians have investigated the properties of trimmed and Winsorized distributions of various shapes, including the stability of means and variances. Strategies for trimming and Winsorizing may be fixed, as when the number of data points to modify is predetermined, or the decision may be tailored to the data in hand (Tukey, 1962). The latter approach is more consistent with the data analyst's philosophy and is illustrated by Daniel and Wood (1971).

The specific data analysis techniques utilized in this study will be discussed in Chapter II. According to the kinds and degrees of usage of the techniques, the firmness of the conclusions drawn from the application of data analysis to curricular problems will vary. Data analysis is not designed to lead to strong inference. Model development, whether data analysis is used or not, requires verification steps to test the generality of the relationships. Curriculum developers, however, are in a position to influence the design of and to coordinate evaluation carried out in local school districts, and opportunities for model verification therefore abound. An evaluation conducted initially in the exploratory mode, yielding less certain conclusions, and followed up by model verification research, promises to serve well the evaluator's purpose of contributing to instructional improvement.

STATEMENT OF THE PROBLEM

The substantive and methodological themes discussed above are represented in the investigation undertaken and reported herein. Two questions guided the present study:

1. Do the variables proposed as measures of curriculum, when used together with descriptors of group baseline performance, serve to predict achievement outcomes?
2. Do data analytic techniques, utilized as a methodology adjunctive to regression analysis, enhance the information yield of an evaluative study?

Illustrative data from two evaluations differing in purpose and subject matter were used to explore these problems. In one study, the effectiveness of a near-completed primary reading program was related to instructional processes in implementing schools, using regression and data analytic techniques, so that implementation could be "tuned" to maximize learning. A second study compared achievement outcomes associated with three published eighth grade mathematics curricula.

The problem of tuning instruction to enhance learning might be more properly considered the province of research on instruction than evaluation *per se*. However, in an era when curricula change rapidly, and new programs are discarded because they fail to yield better pupil performance, this problem is important for the evaluator. Implementation of a new program often is markedly different from school to school and district to district. If the range of instructional practices were represented in an evaluation model, some information about what conditions lead to higher achievement might be revealed. Field tests conducted on products under development typically yield much of the necessary data. In the present study, such data from the Wisconsin Research and Development Center for Cognitive Learning's reading program was utilized. In the 1970-72 school years, 18 Wisconsin and Colorado schools installed and utilized, at the primary level, the Word Attack portion of the Wisconsin Design for Reading Skill Development (Otto & Askov, 1972). Available data include characteristics of the pupil population as well as measures of instructional processes and achievement results.

Data analysis techniques are ideally suited for use in the process of curriculum tuning. Such an activity is long-range, requiring successive studies in which the tentative results from one period of study guide the instructional treatments implemented during the next period. Data analysis in the regression context should help tease out the important variables and yield a richer set of tentative conclusions than might be obtained without its use. However, in the present study, data from only the initial period in such an iterative study were analyzed and interpreted.

To determine how a model with variables representing characteristics of both content and instructional processes might be useful in comparing curricula, eighth-grade mathematics courses in a large junior high school were compared. Over a period of two school years, 1970-1972, three different instructional programs were used, and achievement and instructional process data saved. The texts themselves were analyzed in terms of the content characteristics described in Chapter III.

An experimental design other than that utilized in the present study would be employed, were the average achievement associated with the various curricula the sole focus of the study. As previously mentioned, such studies often fail to yield definitive answers and almost never lead to findings of import beyond the particular study. Outcomes not typically associated with comparative studies may result when the variables in the data set differ in number and kind from those usually employed and when regression and data analysis are substitued for analysis of variance. One goal of the study is to demonstrate the efficacy of the present approach in the context of instructional evaluation.

The data pool in both situations is relatively small considering the fact that the classroom or grade level is the appropriate unit of analysis as discussed in Chapter III. Nevertheless, it is sufficiently large to illustrate data analysis techniques, and the author brings to the analysis process enough acquaintance with the two situations to make the necessary judgments and interpretations.

SUMMARY

The purpose of the present study is to demonstrate the utility of data analysis methodology in evaluative research relating pupil and curriculum variables to pupil achievement. Regression models which account for achievement will result from the application of the methodology to two evaluative problems--one of curriculum comparison and another exploring the relationships between achievement and instructional processes in different schools implementing the same curriculum. Evaluative studies focusing on such questions should yield more information about pupil achievement than evaluations following other models when the following practices are reflected in the design and execution of the study:

1) several dimensions of the curriculum, including material and instructional process aspects, are represented in the set of independent variables;
2) curricular and pupil variables are chosen whenever possible from those conceptualized by educational researchers and known to have a likely relationship to achievement;
3) direct measurements on all variables are incorporated in the data set rather than categorical representations of variables, as in a factorial design;
4) the shape as well as the location of the distributions of pupil achievement before and after instruction is represented in the analyses;
5) the techniques of the data analyst guide the model development process.

These recommendations result from a review of substantive and methodological literature. The first two conditions are essential features of an evaluative research model that contrasts with other extant evaluative models in the nature of its data sets and in its strong linkages with educational research. The third and fourth features of the evaluative prototype under study follow directions set by prominent researchers interested in upgrading the quantitative aspects of both evaluative practice and instructional theory. The fifth facet of the study, the use of data analytic techniques, will contribute to the sparse literature on evaluative research methodology by illustrating data probing techniques that could become part of an evaluator's repertoire.

The development of several models that explain achievement in typical school situations will present opportunities for model verification and refinement. Application of the methodology to a variety of subject matters and student-age groups will ultimately lead to an accumulation of models that suggest the principles of a theory of instruction and the form of their mathematical expression. Model development, in other words, is seen as an important step in building the theory of instruction that Bruner (1964) intended would ultimately guide curriculum evaluation.

Chapter II

REGRESSION AND DATA ANALYSIS

Data analysis is perhaps best thought of as a methodology adjunctive to standard statistical procedures widely used in education and psychology as well as in other disciplines. The data analyst working in the regression fromework uses his data-probing techniques and reveals his philosophy in making judgments before and after the straightforward regression analysis is undertaken. Careful inspection of input data and of residuals alerts the analyst to problems in the data base and to systematic trends not accounted for in the statistical model. The regression equation is evaluated both in terms of its overall adequacy in accounting for the data and in terms of the accuracy of the estimates of the parameters. Knowledge of the consequences of departure from assumptions guides the analyst's interpretation of the equation; discovery of unexpected patterns of relationship between residuals and other variables leads to iteration in the fitting process through the development of alternative models. The statistical methods of fitting and evaluating the model thus are an interplay of classical least squares procedures and newer data analytic techniques.

In actual practice the tedium of considering a single model each time is reduced by considering a class of models utilizing particular subsets of the available independent variables. Preliminary data inspection guides the choice of this subset, and, after the best equation from the class is identified, post hoc procedures shape the next steps or lead to conclusion of the study.

In this chapter, a working knowledge of regression analysis is presumed. First, however, the model will be described and assumptions of least squares estimation reviewed from the data analyst's perspective. Statistical criteria for evaluating the quality of the data set and the goodness of fit are next described, along with the related data analytic tools. Much of the presentation follows the procedures outlined by Draper and Smith (1966), with particular contributions of others noted.

Finally, description of the data analysis sequence makes clear how the interplay of procedures stimulates development and refinement of the initial, tentative model. Chapter IV, in which results are presented, is organized in the sequential pattern of the last section of the present chapter.

THE REGRESSION MODEL AND UNDERLYING ASSUMPTIONS

In fitting an equation, one starts with a statistical model which is "tentatively entertained." This model, $E(Y) = \beta X$, links the response variable Y to the levels of p independent variables, $X_1, X_2, \ldots, X_p$, which determine Y. In practice Y is observed with error, and the equation

$$Y_i = \sum_j \beta_j X_{ij} + e_i \qquad i = 1, 2, \ldots, N, \; j = 1, 2, \ldots, p \qquad (1)$$

accounts for the observations. Least squares, a classical method dating back a century and a half, provides a means of estimating the parameters, β, through minimization of a sum of squares. The observations, Y_i, deviate from the fitted regression, $\hat{Y}_i = \sum_j b_j X_{ij}$, in such a way that sum of the squares of the residuals, $Y_i - \hat{Y}_i$, is a minimum. The least squares estimates are preferred over other means of fitting the parameters because, with certain assumptions, they have the properties of unbiasedness and minimum variance. In other words, the average of the j^{th} b's estimated from samples of a given size drawn from a parent population will equal the population parameters and will also have at least as small a variance as estimators chosen in any other manner. Under the further assumptions of normality of the e_i's, confidence regions for the parameters singly or jointly may be obtained and significance tests performed.

The assumptions underlying regression theory, while not unduly restrictive, are not easily satisfied by field data gathered in the course of educational research. Tukey's (1962) advice that "we need to face up to the necessarily approximate nature of useful results. . . . [inasmuch as] our formal hypotheses and assumptions will never be broad enough to encompass actual situations [p. 61]" provides a perspective from which to review assumptions.

Basic to the use of regression analysis is the assumption that the model being fitted is correct. The properties of unbiasedness and of minimum variance of the estimates depend upon this assumption's being met. Data analytic techniques may be used to check the adequacy of the model several ways following the regression analysis.

The assumption that is perhaps most often disregarded in educational and psychological applications of regression regards the accuracy of the observation of the X's. Regression analysis is founded on an analytic theory that expresses a mathematical, not a probabilistic, relationship between the average value of Y and given values of X. The regression equation itself, while allowing error in the Y's, calls for the X's to be known and not subject to random error. Models of structural relationships which do not require such an assumption about the X's have been proposed and the mathematics outlined (Kendall & Stuart, 1967), but the complexity of the analysis leads the developer of an evaluation model to prefer the regression approach while understanding the consequences of violating the assumption of certainty of the X's.

The problem may be described succinctly: when there is error in X, the correlation between X and Y is attenuated and the value of the parameter estimate is less than would have been obtained were X observed without error. Other perturbations, including impairment of linearity, may also follow.

The remainder of the assumptions underlying the regression model apply to the error term. The assumptions are threefold:

1. The errors are random with mean zero and variance σ^2.
2. The errors are uncorrelated.
3. The errors are normally distributed.

The second assumption is equivalent to the assumption that the observations on Y are uncorrelated. The last assumption, together with the first, is necessary for the most efficient use of the least squares method; obtaining confidence regions and making significance tests depend on it. Since, by the Central Limit theorem, errors in empirical data tend to be distributed normally, the assumption is not overly restrictive. Nonetheless, under the assumption that the model is correct, the observed residuals are pure error, and study of the distribution of these errors provides one test of the adequacy of the model.

Whether assumptions about the model and about e_i are met may be studied empirically after a regression equation has been obtained. The assumption of certainty of X observations is more tractable to investigation through reliability studies independent of the regression analyses. Clearly, if the regression model is used, one might consider attributing any unsatisfactory results to the errors in the X vectors. Furthermore, as Cronbach and Snow (1969) and Suppes (1967) have noted, one should avoid overfitting educational data.

GOODNESS OF THE DATA IN OTHER REGARDS

Strong conclusions require more of the data than simply the meeting of assumption. In particular, it can be shown that certain characteristics of the input data are associated with the variance of the estimates of the parameters. The effects are shown through algebraic substitution in the expression for the variance of the b's in the single parameter case:

$$V(b) = \frac{\sigma^2}{\Sigma_i (X_i - \bar{X})^2} \quad \text{where } i = 1, 2, \ldots, N. \qquad (2)$$

Defining as a scale value m, the deviation root mean square of the X_i's:

$$m = \sqrt{\frac{\Sigma \, (X_i - \bar{X})^2}{N}} \qquad (3)$$

and substituting for the term in the denominator of the expression for $V(b)$,[1] we have:

$$V(b) = \frac{\sigma^2}{Nm^2} . \tag{4}$$

Extension to the two-parameter case involves the correlation coefficient ρ_{12}^2 :

$$V(b_i) = \frac{\sigma^2}{Nm_i^2\ (1 - \rho_{12}^2)} . \tag{5}$$

Equations 4 and 5 show that the variance of the estimate is (1) directly proportional to the variance of the data, (2) inversely proportional to N, (3) inversely proportional to the square of m_i, the scale factor, and (4) in the two-parameter case, directly proportional to the correlation between observations of X variables. From this result we may conclude that experimental error should be well controlled and that it is desirable to have a large N. The third result has implications for the design matrix, while the fourth suggests that X variables should have low correlations with each other. In addition to decreasing the variance of each estimate, under the assumption of normality of the error distribution, attention to these factors makes smaller the area (for $p = 2$), volume (for $p = 3$), or hypervolume (for $p > 3$) of the joint confidence region around the set of parameters.

Further understanding comes from consideration of the joint distribution of X values. The denominator of Equation 2 suggests that $V(b_i)$ is minimized when the values of X are extreme. Intuitively this design is not appealing, and it has been shown by Box and Draper (1959) that, unless the model being tested is correct, a different design is appropriate. Scaling the region of interest of the X's to the interval (-R,R) with $\bar{X} = 0$, they show, in the case where a linear model is fitted but a tendency to curvature exists, that $\Sigma(X_i - \bar{X})^2$ should slightly exceed $\frac{pR}{3}$, where p is the number of X's in the design. In practice the X's are ordinarily distributed throughout the parameter space to approximate a balanced experimental design. Noteworthy is the fact that an extreme data point can unduly influence values of the b_j's.

---

[1] Box, G.E.P. course notes, mimeographed, 1966

Preliminary inspection of the data gives an indication of how well satisfied is each of the desirable data characteristics. The balance in the design matrix, the size of the experimental error, and the correlation between X values need to be known in order to proceed in obtaining a first regression equation or in appraising it.

Balance in the distribution of the X values is revealed either in a graph (for $p = 2$) or in a cross-tabular display (for $p \geq 3$). Ideally the X array is reviewed in light of prespecified definitions of the regions of interest for each X value; these regions were presumably known when the study was planned. In the evaluation situation many pupil and curriculum characteristics are not under the control of the evaluator, and preliminary attention needs to be given to selecting the educational units and materials involved in the study. After data on instructional processes are obtained, another look needs to be taken at the allocation of the X values.

When evidence of major imbalance is found, additional observations at particular points in the parameter space may be sought, or the region of interest curtailed so that an observation on extreme X conditions does not unduly affect the equation. At this point in the analysis, Winsorizing and trimming procedures do not apply, as it is assumed that X's are known without error; the consideration here is either to eliminate or note for later attention a particular data point.

Data inspection further reveals whether there are any observations repeated at the same set of X values. Such duplication makes possible an estimation of pure error through calculating the sum of squares within m repeats, each with n_m observations:

$$SS_{(\text{pure error})} = \sum_{k}\sum_{u}(Y_{ku} - \bar{Y}_k)^2 \quad \text{where } k = (1, 2, \ldots, m), \quad u = (1, 2, \ldots, n_m). \tag{6}$$

Dividing this term by the degrees of freedom, the mean square error for replication is obtained:

$$s_e^2 = \frac{SS_{(\text{pure error})}}{\sum_{i} n_i - k} \tag{7}$$

The denominator here is simply the number of observations taken at repeated X points less the number of X points at which more than one observation was taken. This term can later be compared with the error estimated by the regression analysis, and a test of significance used as one indication of lack of fit.

Finally, the correlation matrix of all X values and observations of Y is usefully inspected prior to the data analysis. Correlations between the X values are noted as well as the magnitude of the correlations between Y and the various X's. A high correlation between two independent variables indicates that both may not be useful in the regression equation, and a decision may be made in advance to select the one that is more easily manipulated or has greater practical significance in some other sense. Correlations between X and Y values are useful in selecting the variable that will enter the regression equation first.

EVALUATING THE REGRESSION EQUATION

The regression equation is evaluated in terms of its overall adequacy in describing the data and in terms of the accuracy of the estimates of the parameters. Statistics ancillary to the regression analysis, such as R^2, are appraised in light of the researcher's goals. The adequacy of the model is checked overall by these statistics and by comparing error estimates obtained in two ways. Assumptions are further checked by techniques of residual analysis originally outlined by Anscombe and Tukey (1963), and problems in fit are diagnosed. In the following sections the procedures used to follow up a regression analysis are described.

Assessing Overall Adequacy of the Model

The adequacy of the model is assessed in terms of the variation in the data unaccounted for by the model. Summary statistics provided in conjunction with computer-run regression analyses are directly interpretable as indices of fit:

1. $100R^2$ is interpreted as the percentage of variation that is accounted for by the regression equation;

2. The square root of the error estimate given by the residual mean square in the analysis of variance table is the standard error of

estimate; this may be interpreted as a proportion of the mean response simply by dividing s by $\bar{Y}$ (Draper & Smith, 1966);

3. The level of significance of the F-value for regression is considered in evaluating the model's adequacy. In this last regard, a student of Box's (Wetz, 1964) has recommended that, in order to ensure the equation's adequacy as a predictive tool at a range of values considerably larger than the standard error of estimate, an F value four times that required for the set level of significance be obtained.

In the iterative situation of the present study it is desirable that criteria for each of these indices be established at the outset of the study. In other words, goals for R^2 and s or s/$\bar{Y}$ are set in advance, as is the α level. Should high standards be met, one might be satisfied that the model was correct--a key assumption. However, explicit study of error in the model further tests this assumption.

Under the assumption that the model is correct, the residual represents only error. When in fact the model is incorrect, the b_j's are biased estimates of the β_j's so that:

$$Y_i = \text{biased function of } X + \text{amount of bias} + \text{error.}$$

In the case of an incorrect model, the residual, $Y_i - \hat{Y}_i$, will be inflated by the bias in the model.

In the preceding section, it was seen that an estimate of pure error can be obtained from observations taken at the same point in the X space. Estimates of pure error and of the error represented by the residuals should be approximately the same if the model is correct. The latter, however, may have a component due to lack of fit, so that the expected mean square for residuals is σ^2 plus a lack of fit factor $L^2/(N-p)$. By partitioning the sum of squares for residuals into an amount for pure error and a remainder due to lack of fit, and an F test is utilized to determine whether the ratio:

$$MS_{(\text{lack of fit})}/MS_{(\text{pure error})}$$

is significant. Table 1 makes clearer these relationships. To get an estimate of pure error presents practical difficulties when the X variables are not under experimental control. Daniel and Wood (1971) have proposed a means of using near neighbors in the X space to get estimates of error. Since their technique ignores unimportant X variables, the problem of locating near neighbors is simplified. Starting with the fitted responses, a measure of squared distance in the "effect space" is calculated for 4N-10 pairs of points whose fitted values of Y are close. The pairwise difference in residuals, Δd, is weighted according to the ordinal position q of the squared distance, and an estimate of error obtained by the formula:

$$s_q = 0.886 \sum_q \Delta_q d/q \qquad q = 1, 2, \ldots, 4N-10.$$

Table 1

ANALYSIS OF VARIANCE TO CHECK LACK OF FIT

Source	SS	df	E(MS)	F ratio
Residual	$\sum_i (Y_i - \hat{Y}_i)^2$	N-p	$\sigma^2 + \frac{L^2}{N-p}$	
Pure error	$\sum_u \sum_k (Y_{ku} - \bar{Y}_k)^2$	$\sum_u n_u - m$	σ^2	
Lack of fit	by subtraction	$N-p-(\sum_u n_u - m)$	$\sigma^2 + \frac{L^2}{N-p-(\sum_u n_u - m)}$	$\frac{MS_{(\text{lack of fit})}}{MS_{(\text{pure error})}}$

The constant 0.886 is the reciprocal of 1.128, and used because the expected value of the difference in selected pairs of 4N-10 observations from a normal distribution is 1.128σ. The estimate of error s_q can be compared directly with the residual root mean square to determine whether the latter overestimates error due to bias in the regression model.

Analysis of Residuals

Further verification of the model and the assumptions upon which it is based is undertaken through analysis of residuals. The examination of errors, incidentally, is not peculiar to regression analysis, and many statistical writers join Anscombe and Tukey (1963) in the opinion that "it is nearly always advisable to calculate the individual residuals . . . , [for] the gain in information of the analyses will usually compensate the effort [p. 144]." Accordingly, even in the case where goals have been met by summary statistics and the test of residual error against pure error was insignificant, assumptions are checked by examining the error *en bloc*. In the event that goals have not been met or any shortcomings in the model have been identified, more extensive analysis of residuals may be of diagnostic use. However, as shall be seen, a particular technique of analyzing residuals is not always uniquely diagnostic.

Distribution of Residuals

The foremost use of residuals comes in studying the pattern of error distribution to determine whether the normality assumption has been satisfied. While standardized residuals may be inspected in list form to check whether about two-thirds of them have an absolute value less than or equal to unity, pattern in the distribution is usually studied graphically. The standardized residuals may be plotted in the form of a histogram to determine whether they are distributed in roughly normal fashion. Alternately, a normal or half normal plot (Daniel, 1959) of the ordered residuals may be inspected. The approximation to normality, of course, depends greatly on sample size, and experience is required to judge normality of residuals from small samples plotted in either way. [Draper and Smith (1966) suggest that experience be gained by plotting standard normal deviates of sample size N; Daniel and Wood (1971) provide several different size sets of standard normal deviates on normal plots.] Lack of normality may suggest that a statistical model other than regression is more appropriate; however, it is not diagnostic of a unique problem. Besides true non-normality of error in the parent population, which would suggest an analytic procedure different from least squares, irregularity in the error pattern may indicate the presence of an outlier. How this possibility should be dealt with is described later.

Residual Plot Against $\hat{Y}_i$

In a second analytic technique, raw residuals may be plotted against the fitted observations, $\hat{Y}_i$. In Figure 1 are depicted the expected dispersion pattern and two kinds of discrepancies that may occur. A relationship between the size of the residual and $\hat{Y}$ (Figure 1.b) indicates that the variance is not constant and that a weighted least squares analysis is required. The nonlinear arrangement in Figure 1.c suggests the need for either a quadratic term or for a transformation on $\hat{Y}$.

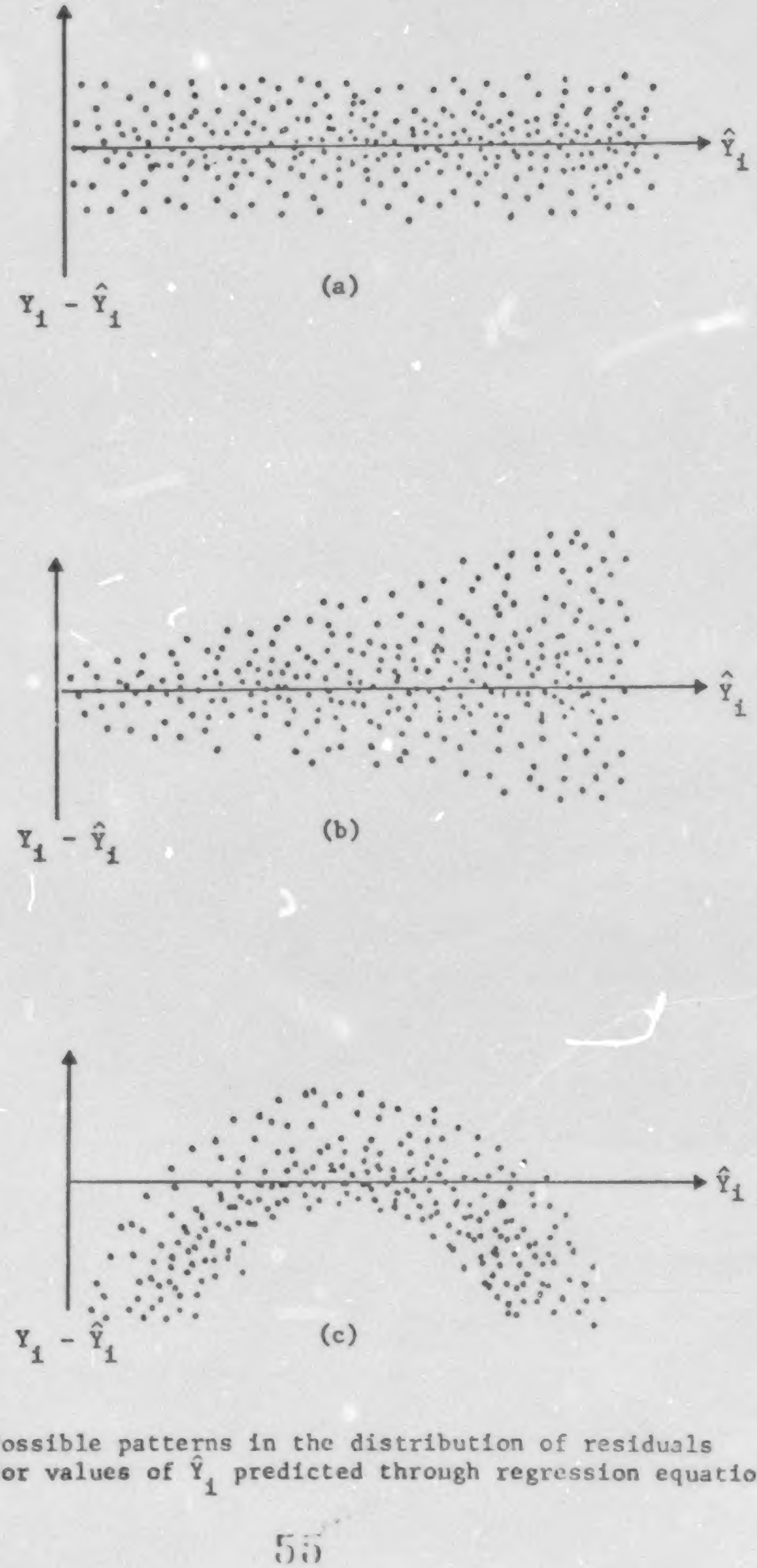


Fig. 1. Possible patterns in the distribution of residuals for values of $\hat{Y}_i$ predicted through regression equations.

Residual Plot Against Terms in X

Similarly, the residuals may be plotted against the independent variables singly or jointly to reveal the same kinds of defects as shown when they are plotted against $\hat{Y}_i$. Discovery of a single X variable for which the pattern in Figure 1.c was observed would suggest that a quadratic term in that variable be added to the equation. The need for cross-product terms may also be identified through residual plots. Draper and Smith see the study of residuals as an important technique for enlarging an inadequate model through identification of quadratic and crossproduct terms.

Residual Plot Against Factors External to the Model

The preceding uses of residuals are sometimes termed internal analyses. A variable external to the defined X space may also be studied in relation to the residuals to determine whether it should be included in the model. In the classic example, data analysts show how time sequence is often related to trends in output in an industrial situation. The opportunities for this kind of residual analysis vary markedly from study to study, and Anscombe and Tukey (1963) regard it as "one of the most important uses of residuals [p. 142]." Both linear and curved trends in the new variable will be revealed by the initial plot.

Outliers

Outliers are revealed in the initial inspection of the error pattern as extreme residuals and, unless dealt with, will persist in aberrant behavior in other residual plots. The first step in dealing with an outlier is to check the data gathering and handling sequence and to determine whether a numerical error has been made that can be easily corrected. Otherwise, the course of action to take is very much a matter of judgment. Examples in the data analytic literature range from Draper and Smith's (1966) caution that "automatic rejection of outliers is not always a very wise procedure [p. 95]" to Daniel and Wood's illustration of iterative data analysis in which, for partly pedagogic reasons, 5 of 21 data points are rejected as mavericks and the remaining data collapsed into 6 data points. Furthermore, there seems to have been little, if any, application of rules for rejecting or modifying data points in education or psychology to provide guidance in the present study.

Outliers in a small data set to be analyzed by regression influence the result in either of two ways. If the bad value is located in the interior of the X space, the constant in the regression equation will be influenced in the direction of its residual. An outlier at an extreme point in the X space affects the estimate of the b_j's. An identified outlier may be rejected or modified. In the latter regard, Charles Winsor proposed that a suspect observation be replaced with the nearest value (Dixon, 1960), a technique now known as Winsorizing the sample. Applied to observations in a regression data set, however, the maverick observation is assigned a new value equal to its original value plus or minus the absolute value of the next largest residual, the sign determined by that of the original residual. The effect is to pull the observation closer to the regression line by compressing its residual

(Anscombe & Tukey, 1963). While the properties of trimmed and Winsorized samples have been examined by mathematical statisticians, no definitive guidelines for the data analyst have resulted. Anscombe and Tukey (1963) suggest that "whereas it may be wise to reject observations whose residuals are very large in magnitude . . ., it may be preferable to modify, but not reject, observations whose residuals are somewhat less large [p. 149]." This intuitively appealing rule simply requires demarcation of the intervals in which residuals are to be treated in either way.

APPLICATION OF PROCEDURES IN THE PRESENT STUDY

In the preceding sections a number of things the data analyst does in the course of a regression analysis have been described. How these processes are juxtaposed and what decision rules are used at each step are to some extent a matter of individual preference. There are, for instance, a number of procedures by which an adequate equation may be secured for several X variables. How outliers are handled also widely varies from one analyst to the next. In this section, the procedures followed in the present study are made more explicit.

The procedural sequence is represented schematically in Figure 2. This summary of the iterative model building process is similar in many respects to that provided by Draper and Smith. However, in the present study, less extensive planning and follow-through steps were taken than are typical in most application, due to limited resources.

Explicit comments with respect to the present study are now directed to each step in the process.

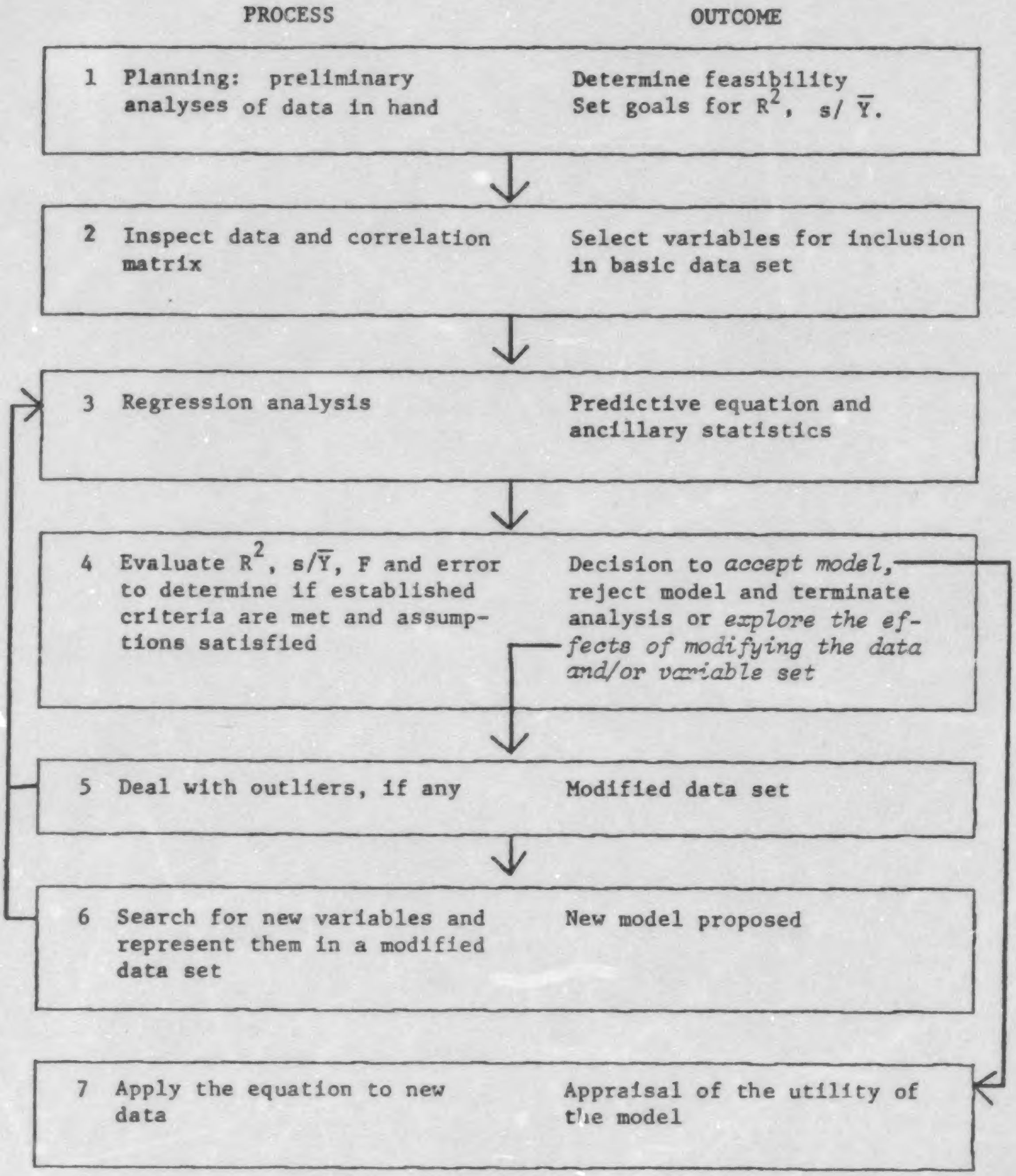
1. Exploratory analysis of available data. Using a subset of the proposed variables, correlations and regression equations were obtained in order to ensure the feasibility of the study and to determine reasonable goals for R^2 and $s/\bar{Y}$.

2. Inspect data and correlation matrix. Of interest are N, the range and distribution of each dependent variable, the joint distribution of data points in the X space, and each r_{X_jY} and $r_{X_jX_{j'}}$. The following points are noted:

a. The balance of data points over the X space is appraised and extreme points flagged so that large residuals associated with them may be given particular attention.

b. Correlations among the X variables larger than 0.60, an arbitrary number, are noted and either (i) one is selected for inclusion in the model if they are conceptually redundant or (ii) both are included and a note is made to take the relationship into account in interpreting either the size of the standard errors of estimate of the associated b_j's or the exclusion of one of these variables from the final equation.

3. Enter basic set of variables in the regression analysis. It is presumed initially that cross product terms and powers of X values will be omitted from the analysis until a line is fitted to the basic data set. Several regression strategies may be used to identify the best equation in the initial run, and they do not converge on a unique equation. A stepwise regression procedure was chosen. Here the variable



PROCESS	OUTCOME
1 Planning: preliminary analyses of data in hand	Determine feasibility Set goals for R^2, $s/\overline{Y}$.
2 Inspect data and correlation matrix	Select variables for inclusion in basic data set
3 Regression analysis	Predictive equation and ancillary statistics
4 Evaluate R^2, $s/\overline{Y}$, F and error to determine if established criteria are met and assumptions satisfied	Decision to *accept model*, reject model and terminate analysis or *explore the effects of modifying the data and/or variable set*
5 Deal with outliers, if any	Modified data set
6 Search for new variables and represent them in a modified data set	New model proposed
7 Apply the equation to new data	Appraisal of the utility of the model

Fig. 2. Sequence of data analysis procedures.

most highly correlated with Y enters the equation first and an additional variable whose partial correlation with Y is highest is inserted if its partial F value meets the set level of significance. As each additional variable is added the contribution of the preceding variables changes. In the stepwise procedure a previously selected variable may be dropped if it does not continue to contribute significantly to the equation. Iteration of the process continues until there are no variables to be added or dropped from the model.

Statistics were provided by STEPREG1, Stepwise Multiple Regression Analysis, a computer program in the STATJOB statistical series (1972) programmed at the Madison Academic Computing Center. The program requires that significance levels be set for including and excluding X variables from the model. In the present study, a significance level of 0.20 was set for the inclusion or exclusion of each variable from the model.

4. Evaluate R^2, $s/\bar{Y}$, and F in terms of criteria. Inspect the significance of the partial F-values associated with each coefficient. Calculate pure error by s_q formula. Make plots of the residual distribution and of raw residuals against $\hat{Y}$. If all criteria are met and the assumptions about the error distribution appear to have been satisfied, the model is accepted as correct. Otherwise, one may decide either that further analysis is warranted or that modification of the data set or variable set may improve the fit.

5. Deal with outliers, if any. Anscombe and Tukey observed that outliers will continue to appear in subsequent residual analyses and suggest that they be handled after the initial analysis. If causes for

large residuals are not identified, observations associated with standardized residuals with an absolute value greater than 3.5 will be rejected, and the residual Winsorizing procedure applied to those whose absolute value lies between 2.5 and 3.5.

6. Search for potential quadratic and interaction terms through plots of residuals against X_j and $X_jX_{j'}$, for $j \neq j'$. Alternatively, search for variables external to the data set. One or more new terms is then added to the variable set.

Iteration of steps 3 through 6 continues until either a satisfactory model is found or no new variables can be identified to improve the equation.

7. Verify the model. As a result of a satisfactory outcome at step 4 new data are utilized to determine whether the model generalizes beyond the data fitted. Some attenuation in R^2 and increase in $s/\bar{Y}$ is expected. Final appraisal of the model overall and the predictive utility of the variables included in it depends greatly upon the verification process, cut short in the present study due to limited resources. In the case of the reading data it was possible to explore the predictive utility of the equations with alternative sets of data. Additional data were not available in the case of the mathematics study, and the data analysis process thus stopped when criteria and assumptions were met at step 4.

Chapter III

VARIABLES IN THE REGRESSION MODEL

In the present study evaluative data are to be fitted to a model that may be represented thus:

pupil achievement = f(pupil traits + curriculum characteristics).

This model, and its statistical expression with interaction terms, employs dependent and independent variables that are similar to those in typical school-based studies of learning and even some laboratory experiments. The experience of educational researchers in using such variables minimizes the problem of identifying measures which might be utilized in the curriculum evaluation model. Still, only a subset of the variables of interest to researchers have practical meaning in a school setting where curricular adoption or modification decisions are often made for the school, class, or some other group of pupils. One problem in selecting or adapting variables for use in a model of curriculum evaluation, then, is to choose from a large number of potential variables those with broad significance in different kinds of school settings. Practical constraints, such as the cost of gathering some kinds of data and school policy regarding the collection or release of certain data, further reduce the possible variables. If exploratory work is to be undertaken in developing integrative models that relate achievement in school settings to pupil and curricular circumstances, however, no option for measuring or otherwise characterizing pupil traits and accomplishments can be regarded as foreclosed.

Despite current controversy about the testing of pupils and interpretation of test scores, the greater challenge in the present study is to explore the curricular variable domain. Under this heading come instructional process variables proposed by curriculum developers and variables representing dimensions of the program content, including its organization. The relationship between achievement and many of the curricular variables has not been studied and may be so tenuous that some possible variables will not warrant a place in a curriculum evaluation model. The present task, however, is one of identifying variables that are sound and subsequently exploring their contribution to the prediction of achievement either singly or in combination with the pupil trait variables.

A model with several independent variables may have terms representing interactions between these variables. Only recently has the interaction term in such a model been regarded as other than a nuisance variable, which, if significant, clouded the interpretation of the main effects. Aptitude-treatment interaction (ATI) research has given new importance to significant interactions, and there are some marginal findings which deserve further investigation. However, the ATI findings apply in the present study only to certain of the interactions between aptitudes and curricular variables, and there is a dearth of research support for the majority of the interaction terms that could be formed by pairwise combination of all independent variables. Again, screening of interaction terms through the statistical procedures will identify those which should be represented in evaluation models.

Pupil characteristics vary on an individual basis while curricular characteristics vary on a classroom, grade level, or school basis. The structure of the curriculum and the teacher's testing procedures, for instance, ordinarily do not vary within a class, even though the pace of instruction may be tailored to each pupil. The appropriate sampling unit for a curriculum evaluation study is thus some collectivity of pupils. Wiley (1970) observed that this fact opens up a number of alternatives with respect to measurement and data analysis. Of particular relevance to the present study are statistics describing the achievement distribution which may be used as both predictor and criterion measures.

In the remainder of this chapter, the particular dependent and independent variables considered for inclusion in a curriculum evaluation model will be reviewed. The unit of analysis problem and the use of statistical descriptors of distributional location and shape will be discussed. Finally, the specific data sets available for the curriculum comparison and curriculum tuning studies will be summarized.

DEPENDENT VARIABLES

The domain of instructional outcomes is commonly divided into cognitive, affective and motor components. Basic research suggests that the explanatory relationships among independent and dependent variables represented in a model will differ between and within these areas (Gagné, 1972). In the present study, the dependent variable of interest is pupil achievement in the cognitive domain and, more particularly, attainment of intellectual skills in reading and mathematics. Furthermore, criterion measures, in the present context, need to represent the outcomes of the instruction provided over a substantial period of time. How overall achievement for some major segment of a program may be best measured is an issue among educators, evaluators, and measurement specialists who argue the merits of norm-referenced and criterion-referenced tests without considering the more fundamental question: What is the nature of cognitive outcomes? Consideration of this question properly focuses initial attention on the population of behaviors from which the content of the test or tests is drawn and defers debate about the technical characteristics or norm-referenced and criterion-referenced tests. In this regard, we shall see that the popular notion that course outcomes are a set of explicitly described and highly specific behaviors poses problems in evaluating the overall program effect.

While the behavioral-objective approach prevails at the present time, two alternative viewpoints about the nature of outcomes should first be mentioned. One school of thought, particularly associated with compensatory preschool education, suggests that an increase in intelligence (or in readiness, an overlapping construct) is a desired outcome (Page, 1972; Wang et al., 1970). The relationship between the content of the curriculum and that of an IQ or readiness test is tenuous, for some tasks in such tests measure behaviors developed mainly outside the classroom. For this reason, and because intelligence testing is not an acceptable practice to some educators (National Education Association, 1972), the use of IQ as a criterion measure is not widespread.

Various organizations of cognitive outcomes are provided by Bloom (1956), Guilford (1967), and Gagné (1972), all of whom classify outcomes into several categories. Distinguishing among these categories might result in less ambiguous description of learning outcomes in a given subject matter and in more accurate prediction of achievement than can be achieved without the use of such organizing constructs. The content of curricula, however, is organized more often topically within a subject matter rather than according to the proposals of educational psychologists, and neither reliable means of classifying content according to the categories nor tests that correspond to these categories are readily available. The gap between curricular practice and learning theory thus currently makes difficult the application of this means of organizing outcomes in an evaluative study.

The approach to describing outcomes that has received widespread support from educators is the use of explicit statements of behavioral objectives which, when attained, are the outcomes of learning. Particular progress has been made during the past decade in identifying and describing carefully the objectives for relatively small segments of a curriculum. Controversy exists, however, on whether or not the identification of specific behavioral objectives facilitates evaluation of overall achievement. At one extreme, Skager (1970) argues that most teachers formulate an instructional plan by selecting, implicitly if not explicitly, specific objectives from some hypothetical population of possible objectives. This position is manifest in the existence and use of banks of objectives from which particular objectives may be drawn by teachers and other educational professionals. The implication of Skager's argument is that the outcome of a course of instruction is a possibly fragmented set of specific behaviors some or all of which should be individually measured to evaluate course effect. Lohnes (1972a) counters with the statement that, while specific objectives are needed for curriculum development, "achievement traits are parsimonious integrations needed in telling about a curriculum [p. 15]." Krathwohl's (1965) conception of several levels of objectives supports Lohnes' argument and makes clear the hierarchical relationships and functional use of objectives at each level. Curricula associated with Project PLAN (Flanagan et al., 1971) and the Wisconsin Design for Reading Skill Development (Otto & Askov, 1972) illustrate how objectives at one level overarch more specific objectives at a lower level. Even so, the two curricula mentioned differ in the granularity of both higher and lower level objectives, and therefore in the approximation to the desired parsimony that is represented by there being few pyramids with an apex representing a terminal behavior. The opportunity presented to the evaluator of such an ideal program is to employ in curriculum appraisal a criterion-referenced test of the most broadly described and highest order behavior at an appropriate time. The concept of such a test (Harris & Stewart, 1971) extends the common meaning and application of criterion-referenced tests but has not been widely used in practice.

Were outcomes organized either according to one of the proposed psychological categories or in terms of one or several parsimonious statements of terminal objectives, clarity and comprehensiveness respectively might characterize evaluative findings. Educational programs needing evaluation are more frequently structured solely in terms of specific behavioral objectives, however. Granularity of objectives, as we have seen, is a relative matter, but inspection of educational programs in use or under development reveals a propensity for objectives that can be attained in the course of a few hours of instruction. The choice that is presented to the curriculum evaluator is to evaluate separately instruction on numerous, highly specific learning objectives or to treat the objectives in the aggregate.

The principal purpose of the evaluation suggests whether the focus should be upon the parts or the whole. When a curriculum is in a developmental stage, there is ample justification to assess the degree of learning that is attributable to particular small segments of the program. When the curriculum is ready for or in wide use, some aggregate measure of performance that is fair to the curricula being evaluated is needed to appraise overall program effect. Measures which are functionally different, such as those used for periodic achievement monitoring and group diagnostic purposes, do not serve this purpose, nor do the cumulative data on program-specific tests provide convincing evidence.

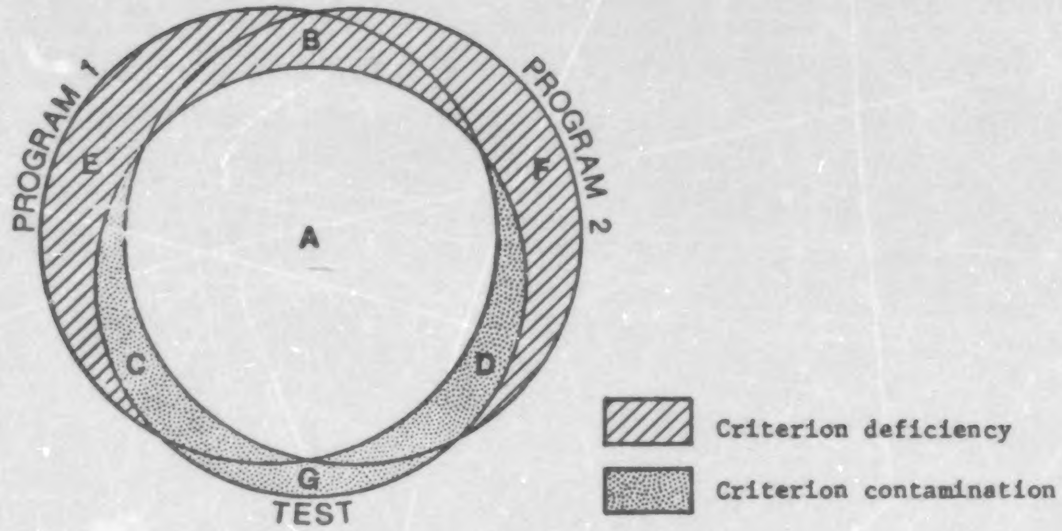
Both norm-referenced, or standardized, tests and combinations of scores from criterion-referenced tests provide aggregate measures. To date the combining of items from tests referenced to specific objectives has been ad hoc (e.g., Skager, 1971). The validity of either kind of test for curriculum evaluation depends upon content definition strategies and upon how closely the test content matches the program content. Advocates of the objective-based systems approach, which includes the use of criterion-referenced tests, seem to have overlooked the fact that in the construction of norm-referenced tests immediate instructional objectives, either given in or derived from curricula, are used to specify the content dimensions for the standardized test (Lindquist, 1951). Here, the objectives common to curricula in wide use are represented in some crude proportion to the amount of instructional time devoted to them, this proportion often being inferred from the percentage of pages in a textbook devoted to an objective (Vaughn, 1951). In ad hoc schemes of forming aggregates of criterion-referenced items, an attempt is similarly made to insure representativeness and balance in the test content. In a standardized test, the content more broadly represents what is taught in all schools, whereas the criterion-referenced aggregate test is ordinarily limited to a single curriculum, perhaps locally defined.

Whether this difference markedly affects the validity of a norm-referenced test for program evaluation of a particular curriculum probably depends upon the degree of consensus among educators about the content of instruction in a given subject matter. Despite assertions to the contrary (Klein, 1971), most items in certain standardized tests of reading and mathematics can, in fact, be keyed to the objectives of a variety of curricula. This is true to a somewhat lesser extent for study skills (Quilling & Wojtal, 1972) and probably for language skills; the correspondence between test items and program content in the social studies and sciences is far more tenuous and suggests that standardized tests may be most useful for subject matters whose content is skill rather than substantively oriented. In terms of a concept attributable to Brogden and Taylor (1950), standardized tests in the skill areas are subject to very little criterion contamination by elements extraneous to a given curriculum. This is a necessary but insufficient test of validity, however, since the inclusiveness and the relative balance of content are also important factors. A standardized test may be less valid if it is biased either by the omission of pertinent content (criterion deficiency) or by distortion of the weighting given to particular content. Examination of tests in relation to the curricula and the judgment of the curriculum specialist are ultimately needed to determine whether a test is valid for a particular curriculum.

While a composite score from criterion-referenced items appropriate to a single curriculum is unlikely to be biased by criterion contamination, criterion deficiency and distortion may decrease its validity for program evaluation purposes. The item sampling techniques that have been proposed for developing composite tests to be used in achievement monitoring (Clausen, 1971; Gorth, 1972) or for program evaluation purposes are presumed to lead to proper representation of the course content; however, the possibility that rather unrepresentative samples of items will be drawn needs to be guarded against through the same kinds of comparison of curriculum and test content as are suggested for standardized tests. It is the author's experience that users of such criterion-referenced aggregates question the validity of particular items for given objectives or of the set of items for a large segment of instruction. These questions suggest that criterion-referenced aggregates and standardized tests have common problems in both face and content validity.

In finding a criterion measure for two or more curricula, the problem of validity is compounded, and it is essential that the degree of commonality between the content of the two programs and the test be large. Lohnes' (1972a) schema in Figure 3, with the key adapted to illustrate Brogden's concepts, shows that for two curricula with overlapping content, a single test may have (1) a relatively large proportion of items common to both contents (Area A), (2) a relatively small proportion both of elements extraneous to both contents (Area G), and (3) unrepresented content common to both programs (Area B). There are two further requirements that a test must meet to be fair to several curricula: first, criterion contamination for one program by the other's content (Areas C and D) should be relatively small and not favor one program; and second, criterion deficiency unique to each program must be small and relatively equal (Areas E and F). In short, the test should be equally valid for each curriculum being evaluated, and Area A should be proportionally large. Such a test is necessary to referee the claims of competing curricula. Procedures for outlining the content of standardized tests guarantee that the content of that test in reading or mathematics will approximate this model; objective banks could probably be used to construct criterion-referenced aggregates fair to two or more curricula, although it is unclear how such an aggregate would differ in form from that of a standardized test.

The preceding discussion is necessary to justify the criterion variables used in the two illustrative evaluations in the present study. The "curriculum tuning" evaluation, that of the Word Attack component


B
PROGRAM 1
PROGRAM 2
E
F
A
C
D
G
TEST
▨ Criterion deficiency
▒ Criterion contamination

A - Content of test and both programs congruent
B - Content of both programs unrepresented in test
C - Content of test extraneous to Program 2
D - Content of test extraneous to Program 1
E - Content of Program 1 unrepresented in test
F - Content of Program 2 unrepresented in test
G - Content of test extraneous to both programs

Fig. 3. Schema of a hypothetical relationship between the content of two programs and a criterion test. (Originally attributed to Cooley and adapted from Lohnes, 1972a, p. 65.)

of the Wisconsin Design for Reading Skill Development (Otto & Askov, 1972), utilized as dependent measures composites of criterion-referenced tests provided in the program materials as well as norm-referenced tests (Table 2). The only criterion measure available in the cross-curricular study of the eighth grade mathematics courses was the Mathematics subtest from the Tests of Academic Progress, Form S, 1971, administered in the course of the school district's system-wide testing program.

In light of the earlier discussion, some comment on the instruments is needed. In the case of the norm-referenced tests used in the reading study, items were cross-referenced to the program objectives and checked for the three kinds of criterion invalidity noted by Brogden. On one test, the Word Analysis subtest of the Cooperative Primary Tests, an excellent fit between program and test content was found; ceiling effects expected at Grades 2 and 3 led to the substitution of the Word Study Skills subtest of the Stanford Achievement Tests at Grade 2, despite its being less valid in terms of Brogden's criteria, and the omission of any standardized test of word attack skills at Grade 3. The mathematics test was not evaluated with respect to the test's validity for the programs being compared. Also, it is noteworthy that the criterion-referenced composite utilized in the reading study and described in Table 2 was formed by sampling subtests rather than items. Because of the age of the group and because different item formats, test directions and sample exercises were associated with each program objective, it was deemed more feasible to construct a composite of several short subtests.

Table 2

CRITERION MEASURES OF PUPIL ACHIEVEMENT
USED IN THE "CURRICULUM TUNING STUDY" OF THE
WISCONSIN DESIGN FOR READING SKILL DEVELOPMENT (WDRSD)

	Criterion	When administered
D_1	Word Analysis subtest of *Cooperative Primary Tests*, Form 13B, 1965	May, Grade 1
D_2	Word Study Skills subtest of Primary I *Stanford Achievement Tests*, Form W, 1966	May, Grade 1
D_3	Composite score from WTRSD[a] criterion-referenced tests of B3 - Beginning Consonant Sounds B5 - Consonant Blends B6 - Rhyming Elements B7 - Short Vowels B10 - Contractions B11 - Base Words & Endings	May, Grade 1
D_4	Word Study Skills subtest of Primary II *Stanford Achievement Tests*, Form W, 1966	May, Grade 2
D_5	Composite score from WTRSD[a] criterion-referenced tests of C3 - Consonant Blends C4 - Long Vowel Sounds C5 - Vowel + *r*, *a* + *l*, *a* + *w* C6 - Diphthongs C12 - Consonant Digraphs C16 - Synonyms & Antonyms D2 - Three-letter Consonant Blends	May, Grade 2

(Continued)

Table 2 (Continued)

	Criterion	When administered
D_6	Vocabulary and Comprehension subtests of *Comprehensive Tests of Basic Skills*, Form Q, Level 1, 1968	May, Grade 3
D_7	Composite score from WTRSD[a] criterion-referenced tests of D2 - Three-Letter Consonant Blends D3 - Silent Letters D4 - Syllabication D5 - Accent D7 - Possessives	May, Grade 3

---

[a] *Wisconsin Tests of Reading Skill Development, Preliminary form*, 1970

The criterion tests were administered under standard testing conditions in each school. Through a test-examinee sampling plan, a class-sized group of primary children was assigned to one of the listed tests in May at the conclusion of approximately one and one-half years of Word Attack instruction. In the case of the junior high pupils, testing occurred in October following eighth grade mathematics instruction. Raw scores were used in calculating the summary statistics used as the dependent variables in the reading evaluation, and standard scores were employed in the mathematics study.

INDEPENDENT VARIABLES

The evaluation model developed in the present study is based on the presumption that pupil and curricular variables contribute to the prediction of achievement. In the educational setting, pupil characteristics account for large amounts of the observed variance in performance and are ordinarily utilized to reduce error in whatever else is being studied (Lohnes, 1972a). In curriculum evaluation, however, attention is properly directed also to curricular variables, because decision making that is guided by the evaluative outcomes is restricted to factors that the educators can control. The model employed in the present study, accordingly, will make use jointly of pupil and curricular variables, while focussing on the latter. These in turn are subdivided into content and instructional process variables.

Pupil Variables

A truism known to theoretical and applied workers in education and other fields is that the best predictor of future performance is past performance (Cook, 1951). Thus, some measure of past educational achievement should relate significantly to the achievement criterion. Some writers, notably Cronbach and Snow (1969) and Lohnes (1972a), argue that intelligence tests scores or "g" factors extracted from a battery of ability tests are better predictors than are academic achievement tests. In the school setting where IQ tests are viewed with reprehension by some, achievement tests serve as good substitutes; the high correlation between aptitude, as measured by intelligence tests, and achievement as measured by standardized tests make the latter an alternative predictor. Socioeconomic status (SES) is also regarded as a good predictor of achievement (Wiley & Bock, 1967), but the correlation of past performance and SES makes the two measures somewhat redundant (Wiley, 1970). Also, the ready availability of achievement scores leads to their choice as covariates in a large proportion of research and evaluation studies. Furthermore, achievement levels of the target pupil groups are considered by educators in making curriculum decisions, whereas variables such as sex and personality traits ordinarily are not. There are thus a number of good reasons why achievement is preferred as a predictor of later performance.

In choosing an appropriate measure of past achievement, the considerations delineated above for the measurement of achievement outcomes hold. Either a norm-referenced test or a composite criterion-referenced measure can be utilized. While one might find higher correlations between the two measures of achievement were the identical test utilized before and after instruction, school achievement tests used in successive years typically have somewhat different, if overlapping contents, and in no case in the present study were the same pretests and posttests utilized for a given age/grade group. The baseline measures available from the evaluative studies, however, are much like the criterion measures: both criterion- and norm-referenced measures were used in the reading evaluation and a standardized test in the mathematics study. The dependent measures for reading are presented in Table 3, while the composite score of two arithmetic subtests in the Iowa Tests of Basic Skills, Form 3, 1964, served as the predictor in the mathematics study.

Curricular Variables

Curricular variables have been classified above into two subsets associate with content and instructional processes. When a single curriculum is evaluated, the instructional process variables are of interest, inasmuch as the content is constant for all field test participants. In cross-curriculum evaluation, however, there may be differences in content as well as in instructional processes.

Since each of the illustrative evaluative studies involves an individualized curriculum, proposed organizations of the curricular variable domain for individualized curricula are relevant. One organization pro-

Table 3

PREDICTOR MEASURES OF PUPIL ACHIEVEMENT
USED IN THE "CURRICULUM TUNING STUDY" OF THE
WISCONSIN DESIGN FOR READING SKILL DEVELOPMENT (WDRSD)

	Predictor	When administered	To predict achievement in
P_1	'Composite score from WTRSD[a] criterion-referenced tests of A1 - Rhyming Words A2 - Rhyming Phrases A3 - Matching Shapes A4 - Matching Letters and Numbers A5 --Matching Words and Phrases A7 - Initial Consonants	September, Grade 1	May, Grade 1
P_2	'Word Analysis subtest of *Cooperative Primary Tests*, Form 13B, 1965	Spring, Grade 1	May, Grade 2
P_3	'Word Study Skills subset of Primary I *Stanford Achievement Tests*, Form W, 1966	May, Grade 1	May, Grade 2
P_4	'Composite score from WTRSD[a] criterion-referenced tests of B3 - Beginning Consonant Sounds B5 - Consonant Blends B6 - Rhyming Elements B7 - Short Vowels B10 - Contractions B11 - Base Words and Endings	May, Grade 1	May, Grade 2
P_5	'Word Study Skills subtest of Primary II *Stanford Achievement Tests*, Form W, 1966	May, Grade 2	May, Grade 3
P_6	'Composite score from WTRSD[a] criterion-referenced tests of C3 -- Consonant Blends C4 - Long Vowel Sounds C5 - Vowel + *r*, *a* + *l*, *a* + *w* C6 - Diphthongs C12 - Consonant Digraphs C16 - Synonyms and Antonyms D2 - Three-Letter Consonant Blends	May, Grade 2	May, Grade 3

---

[a] *Wisconsin Tests of Reading Skill Development*, Preliminary form, 1970

posed by Cooley (1971), and relating to Individually Prescribed Instruction in particular, focuses appropriately on instructional process variables. Another schema, developed by DeVault and his colleagues (1973) for the analysis of various individualized curricula, calls for ratings of program organization as well as instructional processes and associated media.

The main categories of descriptors in each schema are as follows:

- Cooley's classroom instructional process variables
  - Testing procedures
  - Prescriptive (or individualized) practices
  - Teacher skills (e.g., reinforcement)
  - Instructional materials actually used
  - Allocation of time
  - Space and its utilization
  - Teacher's knowledge of the curriculum and children

- DeVault et al.'s categories to analyze individual instruction
  - Program Pattern
    - Objectives
    - Sequence
    - Rate
    - Medium
    - Grouping
  - Learner Assessment Procedures
  - Management of Information
  - Management of Instructional Components

For each schema, numerous particular descriptors are required. These are explicitly described in DeVault's schema and are alluded to by Cooley. Both schemata promise to be helpful to evaluators in measuring curricular variables in the future but also present problems in yielding parsimonious descriptors of curriculum characteristics.

Neither of the preceding schemata were available at the outset of the evaluative studies whose data was reanalyzed in the present study.

Nonetheless, four crude descriptors of the content and instructional processes were attainable from direct inspection of either the program materials or evaluators' data. While crude, these descriptors are of the more parsimonious sort which may be useful in an evaluative model relating achievement to prior pupil performance and curricular variables. The first two variables, named "curriculum decision unit size" and "rate adaptiveness," relate to Cooley's allocation of time category and to rate variables within the DeVault group's schema. An aspect of curricular organization, granularity of specific instructional steps, is proposed for comparing program contents, as is a measure of content difficulty. In no sense is the set of variables used to characterize a curriculum in the present study as inclusive as might be desirable. However, the utility of including measurements associated with such variables in a statistical model for analysis of achievement data has not been demonstrated. Exploring this possibility with a few variables is therefore warranted.

Curriculum Decision Unit Size

The substantive elements of a curriculum can be identified at several levels of specificity as Krathwohl (1965), Flanagan et al. (1971), and others have indicated. The chapter in a textbook or some other topical subsection of a curriculum is a relatively gross organizational unit, while the programmed instruction frame is an impractically fine content unit to consider in summative evaluation. In between falls the material usually associated with one specific behavioral objective.

Kinds of content units should perhaps be conceived of as points on a continuum from the smallest element to the curriculum as a whole. Along this continuum, substantive units which form the framework for the teacher's instructional scheduling can be identified. The function of such a unit is to partition the curriculum into portions about which important instructional or evaluative decisions are made: a grade may be assigned; a decision for the group to advance to new content can be made; or, in an individualized program, an individual may be assigned to a particular topic of study and successful completion noted before the next assignment is made.

Associated with each decision unit is some time interval which is roughly related to the amount of content to which the decision is related. Curriculum decision unit size can thus be measured in terms of the number of days devoted to instruction until a key scheduling or evaluative decision is made.

How might the size of a curriculum decision unit affect achievement? First, very small decision units might lessen achievement by requiring that too great an amount of time be devoted to quasi-instructional activities such as testing and pupil management. Large decision units, on the other hand, may constrain the achievement of the above-average pupil who is held back and thereby affect the mean, variance and skew of the achievement distribution.

Within a curriculum, the decision unit size may vary for individuals within the group or be constant for the group. Of interest in the present study, however, is the average size of a curriculum decision unit. Given the length of time spent by individuals on a particular curriculum decision unit, a mean for that unit can be calculated; the mean of means on all decision units is an expression for the average time between key instructional programming and evaluative decisions. It is presumed that this variable, measured in instructional periods, ranges continuously from 0 to about 40, and has a unimodal and probably positively skewed distribution.

Rate Adaptiveness

Within a class, the teacher may adapt the instruction to individuals in several ways. Adjusting the amount of time spent by each individual is a common practice in individualized programs and was represented in Carroll's (1963) model of school learning. His model called for relating the time an individual spent to the individual's achievement, whereas in the present study some expression of variability in rate across individuals is needed.

In DeVault's schema endpoints and quartiles of the achievement distribution yield five descriptors of the group's rate. These statistics of location have advantages of interpretability to educators but are less complete and less parsimonious descriptors than those derived from moments. In any case, in the present study, the only data available on time spent were the minimum and maximum for a given curricular decision unit and consequently a measure was devised using the range of time spent on each curriculum decision unit. The summary measure, f, is an average across curriculum units, for each of which the range is scaled by its midpoint:

$$f = \sum_{i} \left[\frac{\frac{1}{2}\,(t_{max} - t_{min})}{t_{mid}} \right] \Big/ \, w$$

$$= \frac{1}{2w} \sum_{i} \left[\frac{t_{max} - t_{min}}{t_{mid}} \right] \qquad \text{where } i = 1, 2, \ldots, w$$

t_{max} is the maximum time, in number of instructional periods, spent on a given curriculum decision unit by an individual within the group.

t_{min} is the minimum time, in number of instructional periods, spent on a given curriculum decision unit by an individual within the group.

t_{mid} is half the distance between t_{max} and t_{min}.

w is the number of curriculum decision units presented instructionally.

In effect the numerator in the first expression for f denotes the maximum observed deviation from the midpoint of time spent by the class scaled by the midpoint of the range; this quantity is a proportional descriptor of the time that may be added to or subtracted from the midpoint of time spent on a given unit by individuals in the class. The value of f is an average of these proportional deviations and ranges from 0, when there is no variation, to 1 when the range is twice its midpoint. An example will clarify the interpretation: for a curriculum decision unit on which individuals spent 10 to 20 instructional periods, the value of f is .33, indicating that individuals could vary by as much as one-third from the

15-period midpoint. The distribution of f is probably unimodal with positive skew.

Structural Granularity of the Content

Within curriculum decision units, the content is usually organized into chunks of various sizes. Each of which is associated with some instructional step leading toward completion of the unit. Several such structural constructs are utilized in ATI research, and a few near significant findings suggest that structure, in certain senses, may interact with ability (Cronbach & Snow, 1969). With some consistency, it has been found that less structure leads to somewhat higher attainment on the part of students of high ability, whereas the opposite is true for those with low ability. Curriculum granularity may be measured by the average number of instructional days spent on each instructional step. Equivalent measures are obtained calculating the average time within or across curriculum decision units. The distribution of the measure, r, is bounded on the left by 0, is right-skewed and has typical values between .5 and 3.0.

Content Difficulty

One aspect of the debate among educators over the value of specific behavioral objectives focuses on whether higher order behaviors (Bloom, 1956; Sanders, 1966) are, indeed, represented in curricula completely defined by objectives. This issue has resulted in the need to scale the content of curricula, and Walbesser (1972) has proposed a measure of "puissance." Here the conceptual and performance elements of an instructional step are differentiated in terms of categories provided by Walbesser; the cells in the matrix of combinations of the two kinds of descriptors are weighted differentially, and an average weighting for the content of a curriculum can be determined. This number is low when the pupil produces verbal chains and high when he applies principles to solve a problem. The maximum value of puissance, p, is 21, and Walbesser has suggested that an average value of $p < 6$ is low, $6 < p < 14$ is moderate, and $14 < p < 21$ is high.

Several difficulties are encountered in using Walbesser's index. First, it is apparent that an initial response to a specific kind of problem, such as a number fact, may require more complex behavior than would a later, learned response to the same problem. The time at which a pupil or the majority of a class make the transition from the learning state to the learned state is difficult to identify. Second, the instructional approach appears to make a difference in the rating of a task; different teaching strategies lead to differential ratings both in the conceptual complexity categories and in the performance class. For instance, a child could learn the same fact by rote or by discovery with the values assigned to the performance 1 and 21, respectively. Such a difference is partially attributable to teacher rather than to text and confounds the instructional and content variable categories used in the present study. Therefore, content from the various printed materials was sampled and the textual presentation per se evaluated to determine the level of difficulty of the content. In particular, for the eighth grade mathematics evaluation, 50 tasks in each set of pupil materials were sampled randomly without replacement and evaluated in terms of Walbesser's matrix.

Interactions

Crossproduct measures may be obtained from all possible pairwise combinations of independent variables to expand the data set entered into the analysis for each evaluative study. One of the possible interactions is of theoretical interest, however, and deserves discussion. Considerable ATI research has focused on the differential effects on achievement of content structure. The curriculum structural variable, r, associated with granularity of the instructional steps within a curricular decision unit, represents in the present study an opportunity to investigate the achievement-granularity interaction in a school-based learning situation.

Data analysis procedures may identify additional interactions that improve the achievement prediction equations.

THE UNIT OF ANALYSIS

A curriculum is ordinarily adopted and implemented for some collectivity of pupils within or across schools. This is true even for individualized curricula, such as those associated with Individually Guided Education (IGE), Individually Prescribed Instruction (IPI), and Project PLAN. Typically the subject matter component of such a program is adopted for an entire school or set of schools. The fact that programming decisions are made on an individual or small group basis does not alter the fact that all or most children work within the structure of the adopted program, engaging in similar learning activities, using many of the same materials, and being exposed to the instruction processes favored by one or more teachers. In short, though instruction may be tailored to the pupil, his achievement is partially dependent on classroom or other group factors. Thus, in most educational settings, measurements on classroom groups are the smallest unit on which measurements can be justified in terms of the statistical requirement that observations be independent (Lohnes, 1972b).

For the eighth grade mathematics evaluative study, classroom groups were an appropriate unit of analysis, as the school was conventionally organized. In schools in which classroom divisions have been largely eliminated, a larger cohort of pupils forms the sampling unit. This occurs, for instance, in the multiunit school organization associated with IGE, in which 100 or 150 pupils from two or more grade levels are assigned to a "unit" staffed by a team of teachers (Klausmeier et al., 1971). Instruction with an IGE curriculum such as the Wisconsin Design for Reading Skill Development (WDRSD) is carried out by the team of teachers according to arrangements agreed upon among them, and over a period of time each pupil typically works with all teachers. The conditions of instruction thus are similar for all pupils assigned to the unit. In such an educational setting, the statistical sampling unit should be the teaching unit. Measurement conventions pose a problem in utilizing, as a unit of analysis, a statistic summarizing performance of all pupils in the multi-age organizational unit, however, for the interpretation of standardized test scores depends upon the use of tests appropriate for a given grade level. Accordingly, the multi-age groupings in the multiunit school are usually subdivided along grade level ages for purposes of district-designated achievement testings. This convention was followed in the field test of the Word Attack program of the WDRSD, so that the sampling unit associated with any given measure of the grade level cohort within a unit. Testing arrangements, as mentioned above, called for random selection of pupils from a grade level to form a class-sized group whose performance is measured.

STATISTICAL DESCRIPTORS OF CLASSROOM AND GRADE LEVEL GROUPS

Use of statistical descriptors of group performance is not inconsistent with the philosophy of individualized instruction, if the shape of the score distribution as well as central tendency is taken into account. However, the ubiquitous use of the mean as the sole group descriptor has apparently hindered thoughtful analysis of how distributions might differ as curricula factors are varied. The use of concepts of variance and skew may be useful in describing achievement outcomes associated with various educational policies and practices. In a self-contained classroom without individualization of instruction, for instance, a ceiling effect may be observed in the achievement distribution described statistically, this distribution would have less variance of mean negative skew than would one associated with a program in which means could be individually placed in a curriculum sequence and the rate of learning accelerated for more able pupils. Similarly, an accountability approach toward group achievement in which a minimum level of acceptable performance is identified can be described statistically as an attempt to overcome negative skew by curtailing the left hand or negative tail of the achievement distribution. The Philadelphia reading goal that called for all pupils to read at a 5.5 grade equivalent before they left Grade 6 (Shedd, 1971) is an example of such an approach to accountability; of interest also is how means and variances are affected by the focusing of human and material resources on that subset of the pupil population whose performance relative to the goal is uncertain.

Statisticians typically use moments or cumulants to describe distributions. In the present study, distributions will be described in terms of sample statistics associated with the first three moments, properties of a distribution that are observable in data displays and appear to have significance to educators; the fourth moment, kurtosis, is less interpretable in terms of the shape of the distribution (Kendall & Stuart, 1963) and at least one researcher utilizing it as both predictor and criterion variable concluded that it was "not particularly useful and could be discarded in the interests of parsimony [Lohnes, 1972b, p. 553]."

The mean and sample variance, s^2, are well known measures giving unbiased estimates of the parameters of a hypothetical population of scores that could have been generated by the collectivity of pupils. They are also the first two k-statistics in the family of symmetric functions proposed by R. A. Fisher (1928) as an algebraically simpler means of characterizing distributions. Here the mean, m, and the standard deviation, s, are used as measures of predictor and criterion achievement of pupil groups.

The third k-statistic, k_3, is used to calculate a sample skew coefficient whose sign is negative when the left-hand tail of the distribution is elongated and positive when the reverse occurs. k-statistics are derived from the sums of powers of the deviations from the mean, and therefore values of k_3 for nonsymmetric distributions may be large in absolute value. A more tractable index of skew, g_1, scales k_3 by a function of the sample variance:

$$g_1 = k_3 \;/\; [k_2\sqrt{k_2}]$$

and is used in this study.

Cooley (1971) and Lohnes (1972a, 1972b) have entered moment-statistics and k-statistics respectively into canonical correlations between classroom achievement prior to and following a specified treatment and observed prominent loadings for s^2 and g_1 factors, despite the fact that large sampling errors in k_3 were suspected for the small, class-sized samples. Moderate negative correlations between mean and skew coefficient were observed, and the informational value of g_1 was further verified by the fact that it combined with means to produce a strong canonical regression. Comparison of canonical regressions with multiple regressions predicting distributional characteristics separately led Lohnes (1972b) to conclude however, that in some cases "it would be practical to predict the criterion [mean] and [standard deviation] separately from their regressions on some of the descriptors . . . rather than to bother with canonical criterion factors [p. 553]."

The interpretability of the results in terms of educational practice would appear to be enhanced by not predicting several distributional characteristics simultaneously. In the present study, each of the three distributional indices are predicted singly, although the mean (m), standard deviation (s), and skew (g_1) on the achievement predictor may enter the equation for any criterion. Distributions of scores were obtained only for pupil achievement variables, and the preceding discussion does not apply to the curricular variable domain.

DATA SETS ENTERED INTO THE STATISTICAL ANALYSES

In Chapter I, a twofold application of the general evaluative model was foreseen. First, in a program evaluation context, variation in instructional processes was to be related to achievement outcomes so that implementation could be tuned to maximize performance. Second, several curricula differing somewhat in content were to be compared. The Word Attack component of the Wisconsin Design for Reading Skill Development is associated with the former application, and three eighth-grade mathematics curricula with the latter. While both applications utilize the previously discussed descriptors of group achievement as independent and dependent measures, the curricular variables associated with the two evaluations differ. The particular observations taken on a grade level cohort or classroom group are described in Table 4, where the codes for dependent (D) and predictor (P) measures of achievement refer to Tables 2 and 3 respectively.

The basic data set for each analysis consists of a set of observations on each variable for each pupil group. Associated with this

Table 4

BASIC SETS OF CRITERION AND PREDICTOR VARIABLES

Study/Cohort	Criterion Variables	Predictor Variables
Curriculum tuning: Class-sized samples of pupils from 18 schools completing Grade 1 in 1972	m:[a] D_1[b], D_2, D_3	m: P_1[c]
	s: D_1, D_2, D_3	s: P_1
	g_1: D_1, D_2, D_3	g_1: P_1
		d: curriculum decision unit size
		f: rate adaptiveness
Class-sized samples of pupils from 18 schools completing Grade 2 in 1972	m: D_4, D_5	m: P_2, P_3, P_4
	s: D_4, D_5	s: P_2, P_3, P_4
	g_1: D_4, D_5	g_1: P_2, P_3, P_4
		d: curriculum decision unit size
		f: rate adaptiveness
Class-sized samples of pupils from 18 schools completing Grade 3 in 1972	m: D_6, D_7	m: P_5, P_6
	s: D_6, D_7	s: P_5, P_6
	g_1: D_6, D_7	g_1: P_5, P_6
		d: curriculum decision unit size
		f: rate adaptiveness

(Continued)

GPO 809-783-7

Table 4 (Continued)

Study/Cohort	Criterion Variables	Predictor Variables
Curriculum comparison: 19-8th grade mathematics classes in 1970-71 and 1971-72 school years	m: TAP arithmetic score[d]	m: ITBS arithmetic score[e]
	s: TAP arithmetic score[d]	s: ITBS arithmetic score[e]
	g_1: TAP arithmetic score[d]	g_1: ITBS arithmetic score[e]
		f: rate adaptiveness
		d: curriculum decision unit size
		r: granularity of the curriculum decision unit
		p: puissance of textual content

---

[a] m, s, and g_1 refer to the mean, standard deviation and skew index, respectively.

[b] See Table 2.

[c] See Table 3.

[d] See page 61.

[e] See page 66.

basic set are an ancillary set of crossproduct and squared terms for the independent variables. The methodological procedures described in the preceding chapter were applied to each data set separately; not all variables are represented in the final equation.

Chapter IV

DEVELOPMENT OF THE PREDICTIVE EQUATIONS

The possible contribution of data analysis techniques to the methodology of educational evaluation were explored through the fitting of equations to data from two evaluative studies. In the case of both a curriculum tuning and a curriculum comparison evaluation, equations were developed to predict the mean (m), standard deviation (s), and skew index (g_1) of the distribution of scores on the achievement criterion. One set of equations, that associated with the tuning or "within curriculum" study, had criterion-referenced measures as variables, while both predictor and criterion in the curriculum comparison study were norm-referenced measures.

The model development process, including data analysis procedures, was summarized in Chapter II (Figure 2). First, exploratory work is undertaken to determine the feasibility of building predictive equations on the kinds of data collected, and to set statistical goals for the equations. The results of this initial step are reported in the first section of the chapter. Next, the fitting of equations to data proceeds iteratively until the goals are met or until the equation cannot be developed to meet more nearly the established criteria.

The iterative steps of the data analysis process are presented in detail for the development of the first equation, that predicting mean achievement of third-grade pupils on a criterion-referenced measure. For the remainder of the equations, the principal results only are described.

FEASIBILITY STUDY

Using data in hand, preliminary regression analyses were performed to determine the feasibility of fitting equations to evaluative data. Three questions guided the feasibility study: (1) Do curricular variables of the kind proposed enter into regression equations and enhance the precision of prediction? (2) Does the form of equations with criterion-referenced measures as variables look different from that of equations using norm-referenced measures? (3) What statistical goals should be set for the more carefully developed equations, given the preliminary results?

Tentative answers to these questions were sought; more definite answers were anticipated in the following, more thorough study.

Preliminary Regression Analyses

Regression equations were developed to predict the Grade 2 reading results (dependent variables D_4 and D_5), utilizing various achievement predictors from Grade 1, and two of the proposed curricular variables--curriculum decision unit size and rate adaptiveness. Most of the possible combinations of dependent and independent variables for the Grade 2 cohort were entered in regression analyses to predict the mean and either standard deviation or skew. The equations all included two curricular variable measures; additionally, in the set of equations various combinations of criterion-referenced and norm-referenced tests were utilized as predictor and criterion variables. Data analysis procedures were not used in this study.

In Table 5 some results of the regression analyses are presented. The R^2 values vary from .25 to .85. Given the reasonably large values of R^2 observed in two initial analyses to predict the mean, it may be inferred that equations can be fit which account for much of the variance observed in evaluative data; ancillary data analysis procedures may lead to the development of equations with more terms and higher values of R^2.

Other comments about the results are noteworthy. Both the skew index and the rate adaptiveness measure entered predictive equations, suggesting their efficacy as variables. In equations to predict the mean, high rate adaptiveness and more negative skew were consistently associated with high means. Second, while there was little stability to the particular form of the equation across data sets, there was no suggestion of problems in mixing criterion-referenced and norm-referenced measures. Finally, casual inspection of the data indicated that an outlier or two was present and that a new variable, school organization, might be important.

Statistical Goals for the Study

Statistical criteria for the regression equations can be set in light of the preceding data. Lohnes' (1972b) report on prediction equations using moment descriptors of the reading performance of 219 primary grade classes also contributes useful information in this regard.

Table 5

SUMMARY OF EXPLORATORY ANALYSES

Data Set	Achievement Predictor	Dependent Variable	Terms Entering Equation (Sign. of Coeff.)	Multiple r	R^2
	Criterion-referenced	Criterion-referenced			
A1	P_4: m,s,g_1	D_5: m	m(-),s(+),g_1(-), rate adaptiveness (+)	.92	.85
A2	P_4: m,g_1	D_5: s	g_1(-)	.69	.47
	Norm-referenced	Criterion-referenced			
B1	P_2: m,g_1	D_5: m	m(+)	.79	.62
B2	P_2: m,g_1	D_5: g_1	m(+) rate adaptiveness (-)	.76	.58
	Criterion-referenced	Norm-referenced			
C1	P_4: m,g_1	D_4: m	g_1(-) rate adaptiveness (+)	.66	.44
C2	P_4: m,g_1	D_4: g_1	g_1(+) rate adaptiveness (+)	.66	.44
	Norm-referenced	Norm-referenced			
D1	P_2 : m,g_1	D_4 : m	rate adaptiveness (+)	.50	.25
D2	P_2 : m,g_1	D_4 : g_1	g_1(-) rate adaptiveness (-)	.64	.41

The best R^2 obtained in exploratory study was .85 for data set A1 in Table 5; many lower values were also obtained. Lohnes found that the average R^2 for equations predicting the mean was .77. A goal of R^2 greater than or equal to .80 is a reasonably high expectation when the criterion is the mean and a value between .75 and .80 is satisfactory. Lower standards are required for the prediction of the standard deviation and skew. Lohnes indicated that values of R^2 slipped to .41 and .34 for equations predicting s and g_1, respectively. The values reported for equations A2, B2, C2, and D2 in Table 5 similarly support the establishment of a goal for R^2 not greater than .50 in these cases.

The standard for $s/\overline{Y}$ can be set at a conservative .05 value for the mean. Since most educational tests have less than 100 items and means rarely exceed 60 raw score points, a prediction with 5% error in practice means that the error of estimation involves only one or two raw score points. The goal for the standard deviation is .15 while none is set for g_1 because of its distributional properties. These goals are arbitrary, and experience in the next phase of the study will yield data to evaluate their reasonableness.

As is mentioned in Chapter II, Wetz (1964) has suggested that the F ratio associated with the regression analysis should be about four times its tabled value at the set level of significance, for prediction over the range of the variables to be regarded as satisfactory. Accordingly, a .10 level of significance was set, and the ratio $F_{(observed)}/F_{(.10)}$ will serve as the third criterion in the study. However, it seems unlikely that, with the lower standards set for equations predicting the standard deviation and skew, this criterion can be met.

Additionally, it is desirable that the significance level of the partial F associated with each coefficient be less than or equal to .10. The statement is apparently in contradiction with the earlier statement that a .20 level was set for the inclusion or exclusion of any variable from the model (see p. 48). However, the identification of a variable which contributes weakly to the equation allows the analyst to explore means of improving its contribution, through squaring the term or dealing with outliers, for instance. A .10 level of significance is set for the partial F mainly as a guide in choosing among competing equations.

MODEL DEVELOPMENT IN THE CURRICULUM TUNING EVALUATION

Iterative experimentation and development of models which predict achievement may be useful in refining a curriculum, including associated implementation practices. In the present study, only the first phase of such a process was undertaken. The resulting equations suggest variables which are significant in affecting the achievement distributions; instructional practices might be altered following the initial results, and a follow-up study undertaken. It should, perhaps, be pointed out that all the data analysis procedures outlined in Figure 2 are carried out for a single empirical study as the initial regression equation is built through data analysis techniques, with experimentation deferred until another iteration commences.

In the sections that follow, the data overall are inspected, and equations are then developed separately to predict the mean, standard deviation, and skew coefficient, in accordance with steps 3 to 6 in Figure 2.

Inspection of the Data

Descriptive statistics for the three dependent variables, Y_m, Y_s, and Y_{g_1}, and for the five predictor variables appear in Table 6. These summary data indicate that each variable does, indeed, span a substantial portion of its possible range. Group performance on both the dependent and independent variable aggregates tends both to be high relative to the number of items and to be negatively skewed. The curriculum decision unit size, expressed in number of instructional days, ranges from less

Table 6

DESCRIPTIVE STATISTICS FOR BASIC VARIABLES IN MODEL 1

Variable[a]	Mean	Standard Deviation	Minimum	Maximum Observed	Maximum Possible
Dependent					
Y_m (D_7)	59.98	4.33	50.61	67.20	86
Y_s (D_7)	13.09	2.45	8.08	16.88	
Y_g (D_7)	-.42	.52	-1.69	.41	
Independent					
X_1 (P_6: m)	50.67	5.15	39.67	59.25	66
X_2 (P_6: s)	11.72	2.55	5.75	15.10	
X_3 (P_6: g_1)	-1.00	.68	-2.72	.10	
X_4 (curriculum decision unit size)	9.86	3.57	3.20	15.00	--
X_5 (rate adaptiveness)	.14	.22	0.00	.78	1.00

[a] See Tables 2 and 3 for explanation of the dependent (D) and predictor (P) variables.

than a week to three weeks. While the values for rate adaptiveness are, on the average, low, their range approaches the practical limit.

Correlations between variables are found in Table 7. Here it is apparent that rate adaptiveness is the best single predictor of mean and skew in the criterion measure and will therefore enter the respective regression equations first. The negative relationship between the mean and skew of each achievement measure is noteworthy because of its consistency throughout other data sets listed in Tables 2 and 3. The large negative correlation observed between the standard deviation and mean on the predictor measure is not typical of correlations observed on other data sets and may be attributable to a ceiling effect wherein

Table 7

CORRELATION MATRIX FOR BASIC VARIABLES

	Dependent			Independent				
				Pupil Achievement			Curricular Variables	
	Y_m	Y_s	Y_g	X_1	X_2	X_3	X_4	X_5
Y_m	1.000							
Y_s	-.031	1.000						
Y_g	-.484	-.131	1.000					
X_1	.440	.337	-.378	1.000				
X_2	.496	.237	.183	-.634	1.000			
X_3	-.320	-.577	.357	-.739	.209	1.000		
X_4	.033	.195	-.021	-.040	-.094	.089	1.000	
X_5	.735	-.087	-.428	.524	-.657	-.146	.064	1.000

little spread of scores is associated with a high mean. Rate adaptiveness as an instructional process is positively related to entry performance of the group, as well as to mean criterion performance.

Fitting an Equation to Predict the Mean

A predictive equation was first developed for mean criterion performance. The distribution of data points in the X space is presented in Table 8 for the three best predictors of mean criterion performance--rate adaptiveness (X_5) and the mean (X_1) and standard deviation (X_2) of prior group achievement, the latter two being negatively correlated. The positive correlation between X_1 and X_5, incidentially, is displayed by this tabular presentation. It is apparent that Schools 11 and 18 are near extremes of the X space because of their exceptionally low standard deviations, high means and high rate adaptiveness. In addition, School 15 had the minimum skew coefficient of the set as well as a high mean. The residuals associated with these values were marked for special attention. At this point, it may also be observed that there are no observations on identical points in the X space, although at least Schools 12 and 16 appear to be "near neighbors" on the variables displayed.

First Regression Analysis

In the first regression analysis only rate adaptiveness entered the equation. The regression equation read:

$$\hat{Y}_m = 57.970 + 14.710\ X_5.$$

The standard error of the rate adaptiveness coefficient was 3.39. Results of the associated analysis of variance appear in Table 9. The ratio SS (regression)/SS (total) gave a calculated value of R^2 of .54.

Table 8

ARRANGEMENT OF OBSERVATIONS IN THE X SPACE

		X_5: rate adaptiveness			
X_1	X_2	.00	.01-.15	.15-.30	.30-1.00
38-42	4-8				
	8-12	3 58.5 (.01)			
	12-16				
42-46	4-8				
	8-12				
	12-16	2 56.4 (.82)	9 61.6 (.10)		
46-50	4-8				
	8-12	1 50.612 (-.82)			
		8 59.3 (-.25)			
	12-16	5 60.0 (-.67)	6 57.2 (-.96)	13 61.5 (-.89)	
50-54	4-8				
	8-12		12 55.9 (-.89)		
			16 56.9 (-.80)		
	12-16	7 57.5 (-1.07)		4 61.9 (-.80)	17 67.2 (-1.84)
54-58	4-8				18 57.3 (-.92)
	8-12		10 57.3 (-1.53)		
	12-16	14 60.5 (-1.89)			
		15 63.7 (-2.72)			
58-62	4-8				11 67.0 (-1.30)
	8-12				
	12-16				

Note.--Underscored numbers are school identifiers. Nonparenthesized score is the mean on the dependent variable, D_7; the parenthesized value is the g_1 coefficient for the independent variable.

Table 9

ANALYSIS OF VARIANCE--FIRST REGRESSION

Source	SS	df	MS	F
Total (Corrected)	319.059	17		
Regression	172.405	1	172.405	18.8 ($p<.001$)
Residual	146.654	16	9.166	

The regression root mean square or standard error was 3.03, which is 5.1% of the observed mean. The calculation of s_q, the estimate of pure error from near neighbors, gave a value of 2.93, which is reasonably close to the reported standard error. Although the F value was over four times that required for significance when α is set at .10, neither the values of R^2 nor $s/\overline{Y}_m$ met the established criteria.

Residuals were inspected next. Data in Table 10 indicate that only two schools had large residuals, and that that of School 1 exceeded the bounds arbitrarily set for specific action. Inspection of the data on each of the subtests contributing to the aggregate measure for this school suggested nothing to which the low criterion value could be attributed. Therefore, following the procedure for Winsorizing regression data, the observed value was changed so that its residual was not larger than that for School 15, which had the next largest value. The Y value was thus changed to 52.628. The plot of residuals against $\hat{Y}_{m_i}$ (see Figure 4) was next appraised for checking the assumption of constant variance. The points at $\hat{Y} = 57.97$ represent schools whose rate adaptiveness coefficient was zero. The maverick observation

Table 10

RESIDUAL ANALYSIS FOR FIRST REGRESSION

School	Y-Observed	Y-Computed	Residual	Percent Residual	Standardized Residual
1	50.612	58.362662	-7.750662	-15.31	-2.5601
2	56.433	57.965430	-1.532430	-2.72	-.5062
3	58.523	57.965430	.557570	.95	.1842
4	61.880	60.790191	1.089808	1.76	.3600
5	60.000	57.965430	2.034570	3.39	.6720
6	57.238	58.877593	-1.639593	-2.86	-.5416
7	57.461	57.965430	-.504430	-.88	-.1666
8	59.281	57.965430	1.315570	2.22	.4345
9	61.566	59.804467	1.761533	2.86	.5818
10	57.266	59.186551	-1.920551	-3.35	-.6344
11	67.000	69.396886	-2.396886	-3.58	-.7917
12	55.903	58.701045	-1.839045	-3.23	-.6074
13	61.517	64.174020	3.025980	4.50	.9995
14	60.473	65.321579	1.333421	2.00	.4404
15	63.700	57.965430	5.734570	9.00	1.8941
16	56.862	58.568634	-2.665634	-4.77	-.8805
17	67.200	60.628356	.888644	1.44	.2935
18	66.655	57.965430	2.507570	4.15	.8283

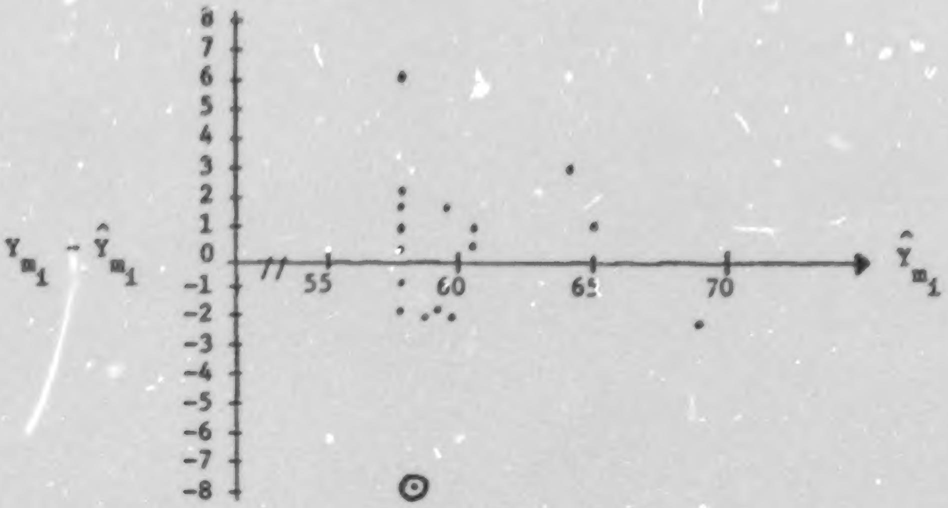


Fig. 4. Plot of residuals against $\hat{Y}_{m_i}$ with X_5 in regression equation.

(circled in the figure) was again revealed, but discounting that point, there appears to be no association between $\hat{Y}_m$ and the absolute value of the residuals.

Regression Analysis with Modified Data

Because of the modified data point, the analysis was repeated using the same variable set. Given the small data set, the coefficients will be slightly altered should the same terms enter the regression equation. Additionally, there is the possibility that different terms will enter the regression equation. Typically the precision of prediction is increased after a data point is Winsorized. On the subsequent analysis, variable X_3 as well as X_5 entered the regression equation:

$$\hat{Y}_m = 56.849 + 13.817\, X_5 - 1.345\, X_3.$$

Computations from the analysis of variance table (Table 11) showed that $R^2 = .63$ and $s/\bar{Y} = .044$, both improved values over the first regression.

Table 11

ANALYSIS OF VARIANCE--MODIFIED DATA

Source	SS	df	MS	F
Total (Corrected)	285.142	17		
Regression	179.967	2	89.984	12.83 (p<.001)
Residual	105.175	15	7.012	

The standard error and s_q error estimates were 2.65 and 2.47, respectively, a wider discrepancy than was observed in the first analysis. Neither R^2 nor the F value met the established criteria.

Inspection of the residual distribution and of the plot against $\hat{Y}_m$, not presented here, was unrevealing. The improvement of the model thus depended upon a relationship between the residual and a second order variable, an interaction or an unsuspected and unincluded variable. The plot of residuals against the predictor mean, X_1, given in Figure 5 is typical of the various plots which suggested no squared or crossproduct terms. Thus, the search turned to external variables.

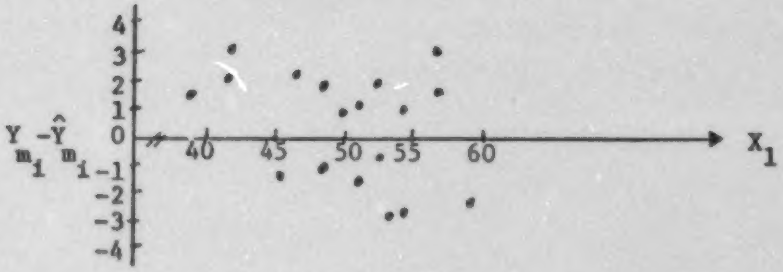


Fig. 5. Plot of residuals against X_1 with X_3 and X_5 in predictive equation.

The one variable proposed by institutional interests and research was school type. Conventionally organized schools may not have adequate flexibility to implement a program as effectively as schools in which staff are organized in teams. Furthermore, the multiunit school (MUS) organization, which provides for teaming of teachers in units, requires gradual introduction of changes over a period of two or more years so that instruction in a multiunit school may not be as effective in recently reorganized schools as in schools that changed earlier. The plot of residuals against the three school types appears in

Figure 6. It is immediately obvious that residuals associated with the old multiunit schools are mostly positive, in contrast to those associated with the other two school types.

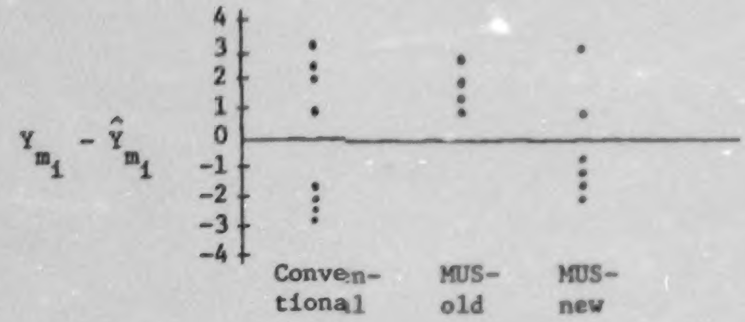


Fig. 6. Plot of residuals against three school types with X_3 and X_5 in predictive equation

Regression Analysis with New Variables

The three school types were expressed as two indicator variables, X_6 and X_7, contrasting, respectively, conventional and multiunit schools and existing and new multiunit schools, and the expanded data set reanalyzed. The contrast between existing and new multiunit schools contributed significantly to the regression by increasing $\hat{Y}_m$ for old multiunit schools and decreasing it for the new schools. The values of the regression coefficients, confidence intervals, and partial F values appear in Table 12. The analysis of variance table for the third analysis is presented in Table 13. With the addition of the indicator variable, X_7, R^2 increased in value to .72 and $s/\overline{Y}_m$ diminished to .040 indicating that knowledge about the school organization clearly improves, the quality of prediction. The criterion for the former value was nearly met, and that for the latter satisfied, though there was no improvement

Table 12

COEFFICIENTS AND RELATED STATISTICS FOR REGRESSION WITH NEW VARIABLES IN THE DATA SET

Term	Regression Coefficient	Standard Error of the Coefficient	90% Confidence Limits Lower/Upper	Partial $F_{(1, 14)}$
Constant	56.338	1.087	54.892/57.784	
X_5	11.184	3.013	7.177/15.191	13.776 $(p<.003)$
X_3	-2.426	1.013	-3.773/-1.079	5.731 $(p<.032)$
X_7	-1.941	.950	-3.205/-.677	4.176 $(p<.061)$

Table 13

ANALYSIS OF VARIANCE WITH ADDITIONAL VARIABLES IN THE DATA SET

Source	SS	df	MS	F
Total (Corrected)	285.142	17		
Regression	204.138	3	68.046	11.76 $(p<.001)$
Residual	81.004	14	5.786	

in F, taking into account its degree of freedom. Furthermore, the residual distribution, Figure 7, looked more normal than in past regressions because of there being small residuals $[\,|Y_{m_i} - Y_{m_i}|<.5]$ not evident in prior analyses.

```
School  -2        -1          0         1         2

 1           *****************.
 2                            *
 3                            .*
 4                          **.
 5                            *
 6                           *.
 7                            .**
 8                            .**********
 9                            .*****************
10             ***************.
11                       *****.
12               *************.
13                            .****
14                            .******
15                            .***********
16                    ********.
17                            .*******
18                            .**
```

Fig. 7. Plot of standardized residuals for regression equation with X_5, X_3, and X_7 entered.

Plots of the residuals against $\hat{Y}_m$ and other included and excluded variables revealed, however, a slight negative correlation with X_1. (See Figure 8.) X_1, in fact, was added fourth to the regression equation in the same analysis that X_7 was added:

$$\hat{Y}_m = 69.840 + 14.160\,X_5 - 4.04\,X_3 - 2.030\,X_7 - .306\,X_1.$$

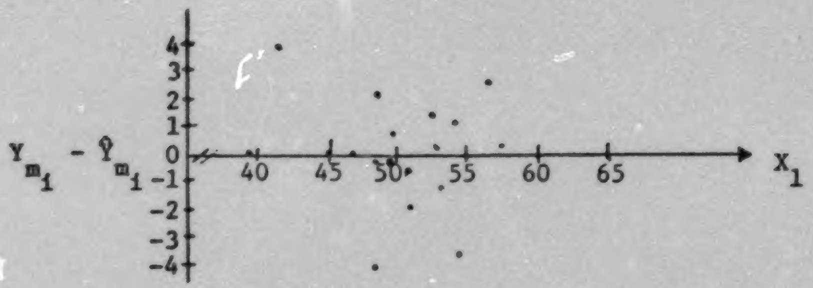


Fig. 8. Plot of residuals against X_1 with X_3, X_5, and X_7 in the equation.

The negative weight given the mean was not intuitively pleasing, and the significance level of its partial F value was large ($p < .164$). Additionally, the changes in $s/\bar{Y}_m$ and R^2 were not impressive. Table 14 provides a summary of the four equations. Although the statistical goals were met by only the last equation, the third equation is better in some respects, including parsimony and the size of the significance levels attached to the partial F's observed for each coefficient. Of practical significance, rate adaptiveness is identified in either equation as a strong predictor of higher achievement, suggesting that this process be stressed in staff development activities and closely monitored during implementation.

Table 14

SUMMARY OF REGRESSION ANALYSES TO PREDICT D_7: m

Equation	Terms Entered	Multiple Correlation Coefficient	R^2	s	$s/\bar{Y}_m$	$F/F_{(.10)}$
1 (original data)	X_5	.74	.54	3.03	.050	6.16
2 (modified data)	X_5, X_3	.79	.63	2.65	.044	4.75
3 (new variables)	X_5, X_3, X_7	.85	.72	2.41	.040	4.67
4 (extension of equation 3)	X_5, X_3, X_7, X_1	.87	.76	2.30	.038	4.16

Model Verification

The last step in the sequence of developing a predictive model through regression analysis is to test the prediction equation on a data set other than that used in developing the equation. A set of data from a standardized test (P_5) was available as a predictor of D_7. Since the skew coefficient was an important predictor in both final equations, and the mean was additionally a predictor in the equation with four terms, it is of interest to see whether the model holds for a different predictor measure. The well known tendency toward skewness for criterion-referenced measures contrasts with the more normal distribution of standardized test scores and raises the question of generalizability of models developed on one set of measures for the other. The model being tested may be represented by the functions:

$$Y = f\ (X_5, X_3', X_7)$$

and

$$Y = f\ (X_5, X_3', X_7, X_1')$$

where the primes indicate that different measurements are used than appear in the data set from which the model was developed. A summary of the correlation coefficients and $s/\ \overline{Y}_m$ proportions is found in Table 15.

Table 15

COMPARISON OF PREDICTIVE EFFICIENCY OF MODEL USING ORIGINAL DATA AND A DATA SET WITH A DIFFERENT PREDICTOR MEASURE OF ACHIEVEMENT

No. Variables in Equation	Achievement Predictor	Multiple Correlation Coefficient	R^2	s	$s/\ \overline{Y}_m$
3	P_6: g_1	.85	.72	2.41	.040
	P_5: g_1	.82	.68	2.70	.045
4	P_6: m, g_1	.87	.76	2.30	.038
	P_5: m, g_1	.71	.51	3.48	.058

The fit of the previously developed equation with three terms to the standardized data is very good, with very slight attenuation in R^2 and small increments of s and s/ $\overline{Y}_m$. The equation which includes the mean as a predictor does not function in a similarly effective fashion, however. This result increases skepticism about its usability, alluded to earlier.

The previous test of the model is in no way adequate, for two predictor variables and the dependent variables were unchanged from the original data set. Several separate, but parallel, data sets were available on which to conduct a more stringent test. Using dependent variable set D_5, predictor set P_4, and the equation with three terms, R^2 is a mere .02; the equation therefore fails to retain its predictive quality. A similar application of the four-variable model gave a residual sum of squares which exceeded the total sum of squares corrected for the mean, indicating that the mean for each school better predicts the average performance of the school than does the predictive equation. The average absolute value of differences between observed values of D_5 and the mean was 4.05 and that between observations and $\hat{Y}$ from the four-variable equation was 5.10, demonstrating clearly the poor quality of the predictive equation in this application. Other applications of both the three- and four-variable equations to new data yielded similarly poor results.

Fitting an Equation to Predict the Standard Deviation

Following the fitting of an equation to predict group mean achievement, a second equation was developed to relate Y_s, the standard deviation of criterion achievement, to the predictor variables. In Table 7, it may be seen that variables in the basic data set, with the exception of predictor skew, had a weak relationship with the criterion measure's standard deviation. The first regression equation for $\hat{Y}_s$ included skew, of course, as well as the predictor standard deviation and curriculum decision unit size. In Table 16, summary information is given. The confidence intervals associated with the coefficients were rather large in some cases. With $R^2 = .55$, derived from the analysis of variance (Table 17), the regression equation accounted for only a slightly larger percentage of the variation than did random error. The standard error of 1.81 is 14% of the mean observed standard deviation.

Table 16

COEFFICIENTS AND RELATED STATISTICS FOR THE FIRST REGRESSION TO PREDICT THE STANDARD DEVIATION

Variable	Regression Coefficient	Standard Error of the Coefficient	90% Confidence Limits Lower/Upper	Partial F (1, 14)
Constant	4.013	2.794	-.907/8.933	2.06 (p<.173)
X_2	.392	.177	.081/.703	4.90 (p<.044)
X_3	-2.473	.662	-3.639/-1.307	13.94 (p<.003)
X_4	.203	.125	-.017/.423	2.65 (p<.126)

The plots against $\hat{Y}_s$ and the independent variables were unrevealing, as illustrated in the plot against X_4 in Figure 9. However, it was found that the average value of residuals for the three school type variables differed somewhat. The mean residual for conventional schools was .94; that for old multiunit schools, -.38; and for new, -.99. The school type contrasts were added as new variables in a second regression

Table 17

ANALYSIS OF VARIANCE--FIRST REGRESSION ON STANDARD DEVIATION

Source	SS	df	MS	F
Total (Corrected)	102.216	17		
Regression	56.319	3	18.793	5.73 ($p = .009$)
Residual	45.898	14	3.278	

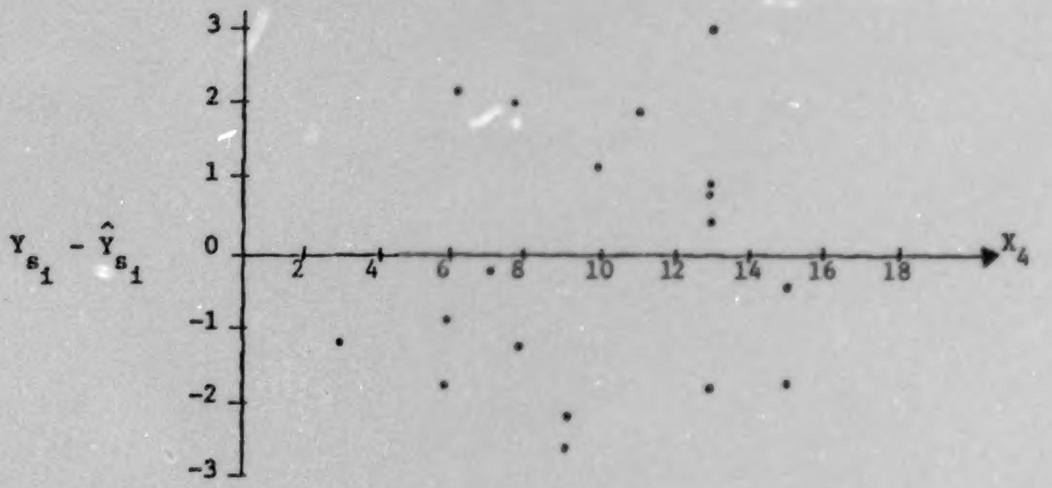


Fig. 9. Plot of residuals against X_4 for first equation for $\hat{Y}_s$.

analysis. Neither the conventional-multiunit school contrast nor the existing-new multiunit contrast was strong enough to enter the regression equation, their partial correlations with the predicted Y_s being .23 and .37, respectively. The original equation was not improved by data analytic techniques; the standard deviation of achievement was predicted with some success by statistics describing the shape of the prior achievement distribution and the size of a curriculum decision unit. The criteria for R^2 and $s/\overline{Y}$ were met, but those for $F/F_{(.10)}$ and the significance level associated with each coefficient were not.

Fitting an Equation to Predict the Skew

First Regression Analysis

Among the basic variables, X_3, predictor skew, as well as X_5, the highest correlate of criterion skew, entered the first regression equation. Values of $R^2 = .27$ and $F_{(2,\ 15)} = 2.80$ suggest that the predictive utility of the equation is quite low.

While the initial residual plots were not diagnostic of problems, new variables were suggested by several plots revealing the behavior of residuals relative to basic variables and interaction terms. First, the residual plot for X_5, presented in Figure 10, showed some curvature, indicating that the square of the rate adaptiveness measure might give a stronger linear relationship with residuals. Schools with zero flexibility scores were less affected by the addition of such a term to the equation; only shifts in the coefficients associated with other variables affect the skew values predicted for such schools.

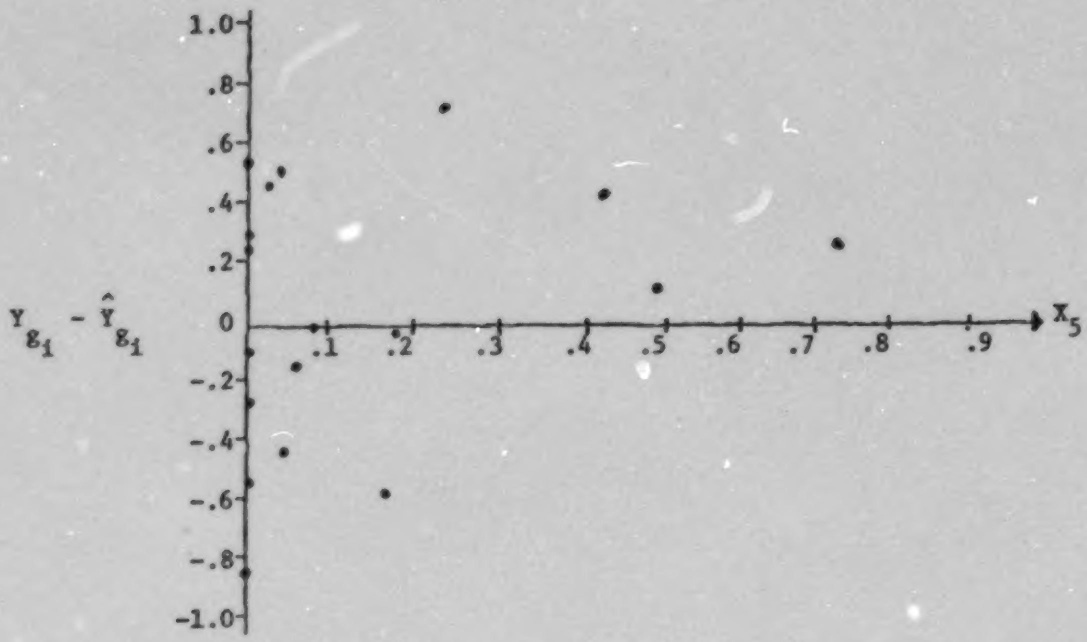


Fig. 10. Plot of residuals from first equation predicting skew against X_5, rate adaptiveness.

A second plot (Figure 11) revealed that the crossproduct of skew and rate adaptiveness was correlated with the residuals. The crossproduct of curriculum decision unit size and rate adaptiveness, X_5, was similarly related to the residuals as Figure 12 shows. These three potential variables were added to the data set.

Regression Analyses with New Variables

In the subsequent regression analysis the skew-rate adaptiveness crossproduct term (X_3X_5) was added first to the regression equation. While R^2 improved, the coefficients of the initial predictors shifted markedly and their confidence intervals widened as comparison of the data in Table 18 will show. It may be recalled that such shifts were not characteristic in the development of equations to predict the mean.

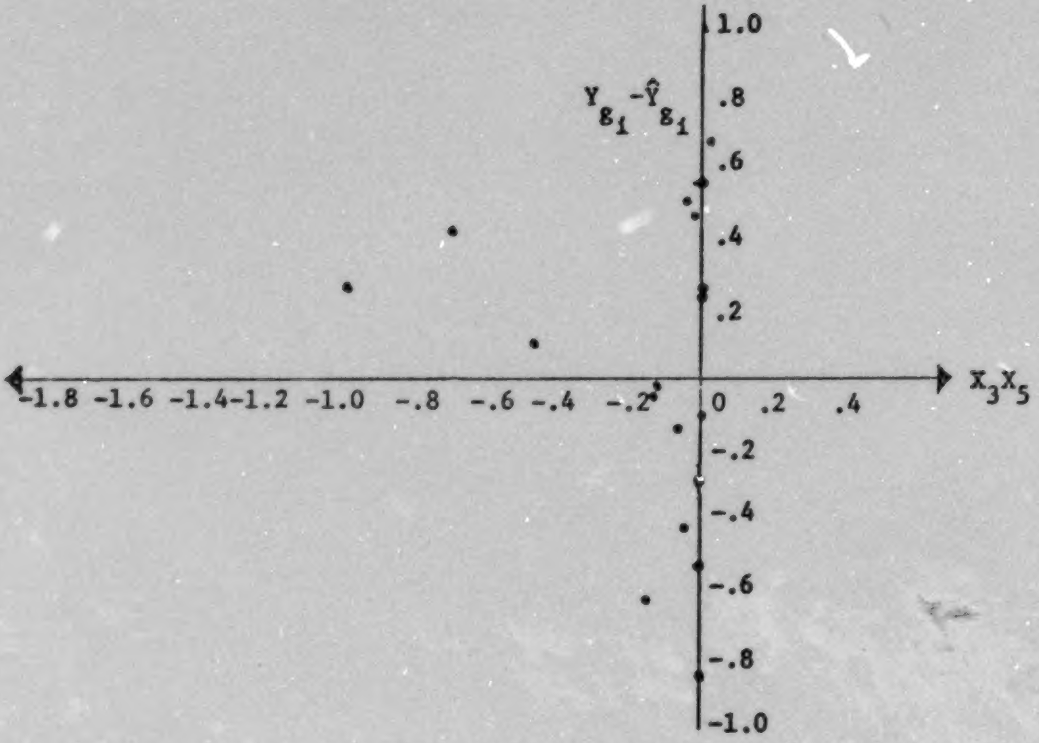


Fig. 11. Plot of residuals against the crossproduct of X_3 and X_5 for the skew predictive equation.

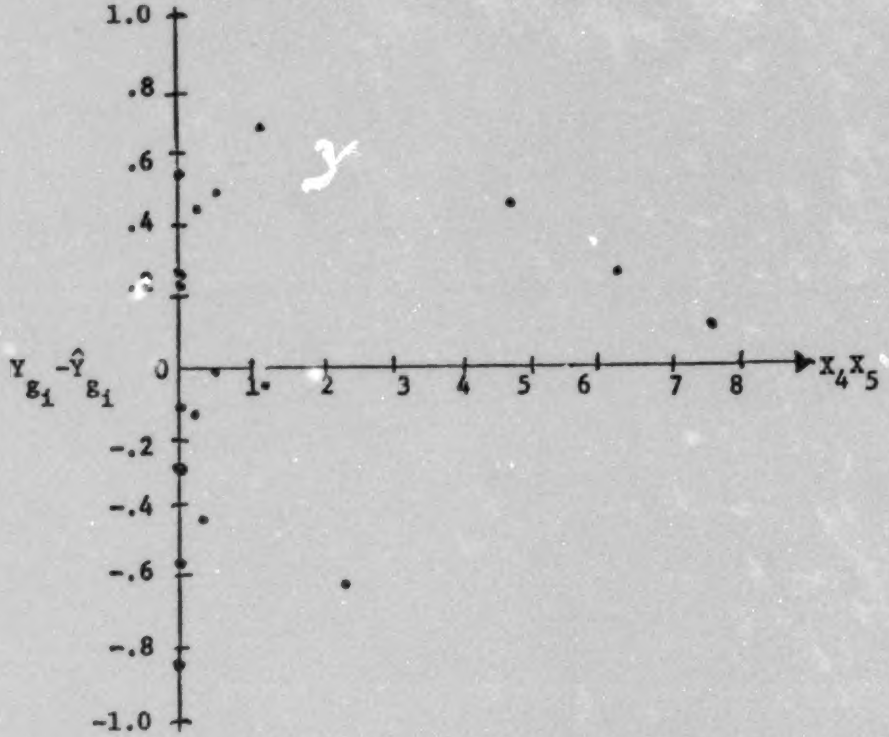


Fig. 12. Plot of residuals from the first equation predicting skew against the X_4X_5 crossproduct.

Table 18

COMPARISON OF ORIGINAL REGRESSION EQUATION AND ONE ADDING A CROSSPRODUCT TERM

Equation	R^2	Terms	Regression Coefficient	Standard Error of the Coefficient	Sig. Level of Partial F value
		Constant	-.069	.206	$p <.744$
1 (basic data set)	.27	X_3	.227	.168	$p <.197$
		X_5	-.915	.530	$p <.105$
		Constant	-.235	.221	$p <.309$
2 (new variable)	.38	X_3	.098	.180	$p <.596$
		X_5	1.943	1.861	$p <.314$
		X_3X_5	2.251	1.411	$p <.133$

The stepwise regression procedure removed X_3 from the equation with negligible attenuation of R^2 and improvement of both the standard error and F value. At this point, the X_4X_5 crossproduct term was added. The significance levels attached to the predictor coefficients were improved as was R^2. Residuals were well behaved in the plots associated with this analysis. The squared flexibility term was redundant with the addition of the new interaction terms. The last two steps are summarized in Table 19.

Finally, the school type variables, which looked promising in a residual plot, were added into the data set, but failed to enter the equation. The best equation to predict skew held X_5 and two of its

Table 19

COMPARISON OF REGRESSION EQUATIONS WITH ONE AND TWO INTERACTION TERMS ADDED

Equation	R^2	Terms	Regression Coefficient	Standard Error of the Coefficient	Sig. Level of Partial F value
		Constant	-.333	.125	$p<.018$
3 (removal of X_3)	.37	X_5	2.347	1.665	$p<.179$
		X_3X_5	2.597	1.228	$p<.052$
		Constant	-.340	.116	$p<.011$
		X_5	5.421	2.294	$p<.034$
4 (addition of X_4X_5)	.49	X_3X_5	3.110	1.178	$p<.019$
		X_4X_5	-.238	.131	$p<.091$

interactions, suggesting the importance of this curricular process in affecting the shape of the distribution. Table 20 presents the final analysis of variance. Though the .49 value observed for R^2 was close

Table 20

ANALYSIS OF VARIANCE--FINAL REGRESSION TO PREDICT SKEW

Source	SS	df	MS	F
Total (Corrected)	4.512	17		
Regression	2.215	3	.738	4.50 ($p<.021$)
Residual	2.297	14	.164	

to the standard, the statistical characteristics of the equation less closely approached the established criteria than did the equation for the standard deviation. Nonetheless, one may safely conclude that rate adaptiveness affects the shape of the achievement distribution.

The Set of Predictive Equations for Reading Achievement

Equations have been fitted to data associated with means, standard deviations, and skew coefficients of group measures of reading performance. The set of best equations is as follows:

$$\hat{Y}_m = 56.338 - 2.426\ X_3 + 11.184\ X_5 - 1.941\ X_7$$
$$\hat{Y}_s = 4.013 + .392\ X_2 - 2.473\ X_3 + .203\ X_4$$
$$\hat{Y}_g = -.340 + 5.421\ X_5 + 3.110\ X_3X_5 - .238\ X_6X_5.$$

Coefficients for skew (X_3) and rate adaptiveness (X_5) appear in more than one equation.

The skew values observed in the predictor achievement measure had negative tendencies; both higher means and standard deviations are predicted for groups with marked left-tailed skew in the score distribution of prior achievement. Rate adaptiveness, observed in over half the schools, was associated with a higher mean and more right-tailed skew in criterion performance.

The final equations are summarized in terms of the established criteria in Table 21. The equations to predict both the mean and standard deviation met some of the established criteria. However, R^2 for the equation predicting Y_m was disappointingly low. The applicability to new data of the equations predicting the criterion standard deviation and skew index was not explored because of the low R^2.

Table 21

REGRESSION EQUATIONS PREDICTING THE DISTRIBUTION OF READING ACHIEVEMENT

Equation	Terms Entered	R^2	s	s/ $\overline{Y}$	$F/F_{(.10)}$
$\hat{Y}_m$	X_3, X_5, X_7	.72	2.41	.04	4.67
$\hat{Y}_s$	X_2, X_3, X_4	.55	1.81	.14	2.27
$\hat{Y}_g$	X_5, X_3X_5, X_4X_5	.49	.41	--	1.79

MODEL DEVELOPMENT IN THE CURRICULUM COMPARISON EVALUATION

Assuming that curricula differ primarily in the degree to which particular characteristics are incorporated rather than in kind, and that the several characteristics are well balanced across curricula, the attributes that produce a stronger effect may be uncovered by a regression approach. Accordingly, three curricula used in an eighth grade mathematics curriculum were analyzed in terms of the curricular variables described in Chapter III and these characteristics related, along with descriptors of group baseline performance, to distributional statistics on the achievement criterion. Data were not available to carry out a verification study.

Inspection of the Data

In the curriculum comparison study, variables included descriptors of group performance on standardized measures as well as all the curricular variables discussed in Chapter III. Two observations regarding the summary data on the variables (Table 22) are warranted: (1) achievement

Table 22

DESCRIPTIVE STATISTICS FOR BASIC SCALED VARIABLES IN THE CURRICULUM COMPARISON STUDY

Variable	Mean	Standard Deviation	Minimum	Maximum
Dependent				
Y_a (TAP: m)	45.26 [a]	4.75	36.55	50.9[illegible]
Y_b (TAP: s)	6.92	1.68	4.14	9.90
Y_c (TAP: g_1)	-.10	.50	-.93	.92
Independent				
X_1 (ITBS: m)	7.62 [b]	.75	6.47	9.06
X_2 (ITBS: s)	1.24	.23	.88	1.74
X_3 (ITBS: g_1)	.12	.54	-.86	1.19
X_4 (curriculum decision unit size)	17.6	.22	11.10	27.70
X_5 (rate adaptiveness)	.13	.23	0.00	.61
X_6 (structural granularity)	2.68	.44	2.34	3.39
X_7 (puissance)	.76	.04	.71	.81

[a] Expressed in raw score units

[b] Expressed in grade equivalent units

score distributions were more normal in this data set than in the last, as indicated by skew index values close to zero, and (2) the additional scaled curricular variables--puissance and structural granularity--did not span a large portion of the possible X space.

The correlation matrix (Table 23) holds a number of large correlation values. Predictor mean and skew are highly correlated with the criterion mean and skew, respectively, and there are as well a number of strong relationships between the curricular variables and the criterion descriptors. Curriculum decision unit size and structural granularity are both inversely related to the group mean and standard deviation. Rate adaptiveness is positively correlated with the standard deviation of achievement. Unexpectedly, there is a strong negative correlation between puissance and class means before and after instruction.

Other relationships among variables are noteworthy. In the case of the achievement predictor measure a negative correlation between the mean and skew index was again observed. More critically, there are strong relationships between the structural granularity measure and both puissance and curriculum decision unit size. While the measures are not conceptually redundant, the .82 correlation between structural granularity and curriculum decision unit size presents statistical redundancy. A case could be made for deleting one variable from further consideration. Furthermore, the fact that only three observations were taken on puissance--one for each curriculum studied--suggests that low confidence be placed in its correlation values.

Table 23

CORRELATION MATRIX FOR BASIC VARIABLES IN THE CURRICULUM COMPARISON STUDY

	Dependent			Independent						
				Pupil Achievement			Curricular Variables			
	Y_a	Y_b	Y_c	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Y_a	1.000									
Y_b	.537	1.000								
Y_c	-.022	-.124	1.000							
X_1	.791	.157	-.072	1.000						
X_2	.387	.344	.171	.093	1.000					
X_3	- 278	.148	.509	-.431	.482	1.000				
X_4	-.615	-.659	.008	-.178	-.359	.064	1.000			
X_5	.203	.617	.053	-.095	.370	.200	-.604	1.000		
X_6	-.864	-.624	.004	-.617	-.178	.277	.822	-.319	1.000	
X_7	-.757	-.331	-.072	-.835	.076	.330	.228	.092	.730	1.000

A plot of observations in the X space for curricular variables (Figure 13) reveals clusters of points associated with the three curricula, indicating that instructional practices were not sufficiently different within curricula to distribute widely the observations on 19 classes. It is noteworthy also that no rate adaptiveness was observed for two curricula. Inspection of records provided by teachers in the study revealed, in fact, departmental consistency with respect to the content selected and allocation of time to that content. All curricular variables are therefore confounded with the three curriculum policies followed by the mathematics department.

Care is required in the interpretation of the contribution of these variables to the regression equation.

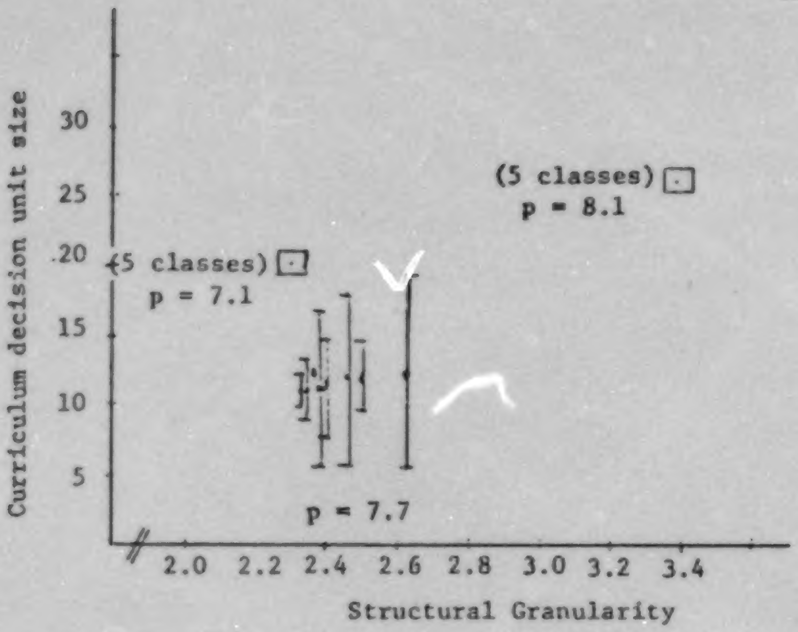


Fig. 13. Plot of X space for curricular variables in the mathematics study. Bars associated with some points represent the degree of rate adaptiveness by showing the range in the average curriculum decision unit size for individuals in the classes or which rate adaptiveness measures were nonzero. Squared points represent sets of classes.

Fitting an Equation to Predict the Mean

First an equation was developed to predict the class mean on the criterion. In the first regression analysis three variables entered in the prediction equation:

$$\hat{Y}_a = 34.577 + .027 X_1 + 5.236 X_2 - 6.028 X_6.$$

The analysis of variance (Table 24) gave $R^2 = .91$ and a standard error of 1.53. A large residual was associated with the class whose criterio performance was lowest. Furthermore, the residual distribution had a number of moderate positive values but, with the exception of the outlier, very small negative values. While this particular distribution

Table 24

ANALYSIS OF VARIANCE--FIRST REGRESSION TO PREDICT MEAN ARITHMETIC PERFORMANCE

Source	SS	df	MS	F
Total (Corrected)	405.569	18		
Regression	370.488	3	123.496	52.80 ($p<.001$)
Residual	35.082	15	2.339	

looked nonnormal, that depicted in the plots associated with the first two steps of the analysis did not, suggesting that the observation was not bad but attributable to the addition of the third term to the equation. There was no discernible trend for residual values to change with $\hat{Y}_a$ nor with any of the included or excluded X values. However, the crossproduct of two of the terms, X_2 and X_6, was negatively related to the residual. This term was added as a variable and did, indeed, load in the equation, improving the fit and substantially reducing the residual previously identified as an outlier. The residual pattern (Figure 14) looked more normal, and external variables such as curriculum and teacher were, by inspection, determined to be unrelated to the residuals.

In Table 25, the first and second prediction equations are summarized. In both cases R^2 exceeds .90, and $s/\overline{Y}_a \leq .02$. All criteria are met by both equations; except for the poor behavior of residuals on the simpler equation, it might be preferred in the interests of parsimony.

```
Class                  Standardized Residuals                   Teacher   Curriculum
          -2           -1           0           1            2
1                                   .********                      1            a
2                                   .***********                   1            a
3                                  *.                              1            a
4                         **********.                              2            a
5                             ******.                              2            a
6                                   .**************                2            a
7                          *********.                              3            a
8                                 **.                              3            a
9                               ****.                              3            a
10                              ****.                              1            b
11                                  .******                        1            c
12                                **.                              1            c
13                                  .**************                1            c
14                               ***.                              1            b
15                   ***************.                              3            c
16                                  .****                          3            b
17                           *******.                              3            b
18                                  .*************                 3            b
19                           *******.                              3            c
```

Fig. 14. Plot of residuals for 19 classes after terms in X_1, X_2, X_6, and X_2X_6 enter the equation to predict mean class achievement.

Table 25

DATA FOR TWO EQUATIONS WHICH PREDICT THE CLASS MEAN IN ARITHMETIC ACHIEVEMENT

Equation	Terms Entered	Regression Coefficients	Standard Error	R^2	s	$s/\bar{Y}_a$	$F/F_{(.10)}$
1	Constant	34.577	7.202	.91	1.53	.020	21.2
	X_1	.027	.006				
	X_2	5.236	1.627				
	X_6	-6.028	1.056				
2	Constant	-4.646	16.842	.94	1.32	.017	23.05
	X_1	.028	.005				
	X_2	37.094	12.797				
	X_6	9.007	6.071				
	X_2X_6	-12.649	5.050				

Fitting an Equation to Predict the Standard Deviation

The best predictor of the group standard deviation was curriculum decision unit size, this variable being negatively related to the criterion. The first prediction equation additionally held terms in X_4 and X_6:

$$\hat{Y}_b = 18.134 - 1.335\ X_4 - .838\ X_5 + 3.405\ X_6.$$

It may be recalled that these variables are badly confounded with the curricula being compared. Nonetheless, the residuals were well distributed and the equation accounted for 59% of the observed variation in criterion standard deviation. Residual plots were suggestive only of testing the contribution of the X_2X_5 crossproduct, and this term was added to the model in a second analysis with R^2 increasing to .67. The X_7 term was subsequently dropped in the stepwise regression process. A summary of the initial and subsequent predictive equations appears in Table 26. Equation 3, accounting for 63% of the observed variance in the criterion, very nearly meets the s/ $\overline{Y}$ criterion of .15; the ratio of the observed F to $F_{(.10)}$ is 3.45.

Table 26

DESCRIPTIVE INFORMATION REGARDING EQUATIONS SUCCESSIVELY DEVELOPED TO PREDICT CLASS STANDARD DEVIATIONS IN ARITHMETIC

Equation	Terms Entered	Regression Coefficient	Standard Error	R^2	s	s/ $\overline{Y}_b$	$F/F_{(.10)}$
1	Constant	18.134	5.753	.59	1.18	.17	2.86
	X_4	-.838	.589				
	X_5	3.405	1.639				
	X_7	-1.335	.795				
2	Constant	18.093	5.319	.67	1.09	.16	2.98
	X_4	-.865	.545				
	X_5	-8.048	6.269				
	X_7	-1.323	.735				
	X_2X_5	8.335	4.427				
3	Constant	18.379	5.428	.63	1.11	.16	3.45
	X_4	-1.399	.748				
	X_5	-.738	.547				
	X_2X_5	2.820	1.093				

Fitting an Equation to Predict the Skew

In the first analysis to predict skew the only variable entering the equation was predictor skew. The value of $R^2 = .26$ indicated that this equation accounted for about one-quarter of the observed variance. Also, two large residuals were associated with the maximum and minimum observed skew indices on the dependent measure. It was also apparent from Figure 15 that the direction of the residuals was associated with a teacher variable, the residuals for teacher 1 being predominantly positive. The largest residual was modified in accordance with procedures, a teacher "dummy" variable entered, and the analysis repeated. The teacher contrast, X_8, entered the equation in this analysis. The second largest residual, not modified, was increased by the addition of the teacher contrast term, and the new variable had a nonsignificant partial F value as Table 27 shows. The observed values of $F = 4.43$ and $R^2 = .36$ were unsatisfactory.

Table 27

COEFFICIENTS AND RELATED STATISTICS FOR THE FINAL EQUATION TO PREDICT SKEW

Terms	Regression Coefficient	Standard Error	Partial F (2, 16)
Constant	-.155	.096	2.62 ($p<.126$)
X_3	.414	.184	5.05 ($p<.040$)
X_8	-.144	.105	1.90 ($p<.187$)

```
 Class           Plot of Standardized Residuals           Teacher  Curriculum

          -2         -1         0         1         2
 1                              .************                1         a
 2                   ***********.                            1         a
 3                              .**                          1         a
 4                       *******.                            2         a
 5                              .***                         2         a
 6                              .****                        2         a
 7                           ***.                            3         a
 8                       *******.                            3         a
 9                           ***.                            3         a
10                              .***                         1         b
11                              .******                      1         c
12                              .***********                 1         c
13          ********************.                            1         c
14                              .************************    1         b
15                    **********.                            3         c
16                              .*                           3         b
17                              .**                          3         b
18                      ********.                            3         b
19                              *                            3         c
```

Fig. 15. Plot of residuals from initial equation to predict skew in the distribution of arithmetic scores, with one variable entered in the regression equation.

The Set of Predictive Equations for Mathematics Achievement

The set of equations developed to predict class performances in the Tests of Academic Progress Arithmetic subtest may be reviewed.

$$\hat{Y}_a = 34.517 + .027\ X_1 + 5.236\ X_2 - 6.028\ X_6$$

or

$$\hat{Y}_a = -4.646 + .028\ X_1 + 37.094\ X_2 + 9.007\ X_6 - 12.649\ X_2X_6$$

$$\hat{Y}_b = 18.379 - 1.399\ X_4 - .738\ X_5 + 2.820\ X_2X_5$$

$$\hat{Y}_c = -.155 + .414\ X_3 - .144\ X_8$$

Predictor mean and skew entered their respective equations, and the curriculum decision unit size variable loaded in each equation for the mean and standard deviation. Statistical criteria were met for both equations predicting the mean, but not for the equations predicting the standard deviation and skew.

Chapter V

CONCLUSIONS AND DISCUSSION

The objectives of the thesis were twofold: (1) to explore the possible contributions of data analysis to evaluative research and (2) to determine whether variables proposed as measures of curriculum, when used together with descriptors of group baseline performance, serve to predict achievement outcomes. Positive findings with respect to the two goals would suggest that the dimensions of the study should be incorporated in a model for evaluative research that has more extensive quantitative attributes and a stronger linkage with other educational research paradigms than do current evaluation models.

The questions of the study were investigated empirically through the analysis of data from two curriculum development field tests. Achievement measurements resulting from use of a single instructional program were related to pupil baseline performance and instructional variables in the first phase of a curriculum tuning study. Next, the comparative performance of class groups using three curricula was investigated in relation to four dimensions of the curriculum and baseline achievement data. The conclusions that were drawn from these studies are presented in the next section. Subsequently, the goals of the thesis are appraised in light of the findings.

AN APPRAISAL OF THE EFFECTIVENESS OF THE TECHNIQUES FOR PURPOSES OF EVALUATION

Curriculum Tuning Study

Equations were fit to data gathered in the field test of the Word Attack area of the *Wisconsin Design for Reading Skill Development* for three reasons: to describe the relationship between criterion performance and the predictor variables in the data set; to identify particular curricular variables important in predicting achievement; and to provide equations whose usefulness in accounting for achievement in other educational situations could be tested.

Inspection of the set of equations that predicted criterion performance revealed the following information:

(1) The mean of a grade level group was predicted by the skew of the distribution of scores on the baseline achievement measure, rate adaptiveness, and school organization. High means were associated with more negative skew, high rate adaptiveness, and an existing flexible school organization.

(2) The standard deviation of group performance was a function of the standard deviation and skew of prior achievement and average size of the curriculum decision unit. A larger baseline standard deviation, greater curriculum decision unit size, and more negative skew were associated with large standard deviations.

(3) The skew index was predicted by rate adaptiveness and two of its interactions. Noteworthy is the importance of the basic rate adaptiveness variable in predicting the shape of the distribution. Within

the range of the predictor skew and curriculum decision unit variables, any degree of rate adaptiveness made the distribution of group performance more right-tailed than it would otherwise have been. This result was intuitively pleasing when coupled with the fact that the mean likewise increased. In short, rate adaptiveness as an instructional practice was associated with a positive shift in the location of the distribution and a tendency toward spread in the achievement of the more able students. This conclusion, however, should be tempered by the observation that rate adaptiveness may be a proxy measure of skillful teaching in other senses.

(4) All three equations behaved statistically in approximately the expected manner, corroborating data provided by Lohnes (1972b). In particular, mean criterion performance was predicted best, and the standard deviation somewhat more accurately than the skew index. Of particular interest was the presence in the predictive equations of distributional statistics other than the mean, and the strong performance of the rate adaptiveness variable. Taken together the results may be interpreted to confirm the convictions of proponents of individualized instruction. This kind of empirical data has been needed for several years to counter the comments of Cronbach and Snow (1969) that individualized instruction is atheoretic and of Gage and Unruh (1967) that positive findings are sparce.

In appraising the value of this information to the consumers of evaluation, it is useful to review the results of the earlier field test analysis. Here the sole finding of import was that, with program implementation, group means at all grade levels were significantly better on the criterion-referenced tests than were those of a comparison group of pupils at the same grade level in the same schools two years earlier (Quilling & Otto, in press). The fact that the better-performing group had experience in taking many of the criterion tests as an integral part of instruction lessened the confidence attached to the positive results.

There is, of course, no reason why the evaluator might not wish to utilize both analysis of variance and the regression approach illustrated here. Each result enhances the other. The experimental-control comparison gives the consumer some confidence that the program leads to higher achievement than do other programs, while the regression equations help explain how the instruction achieves its effects. The methodology used in the curriculum tuning study can therefore be judged to meet the goal of enhancing the information yield of an evaluation.

Despite these positive outcomes, the model verification test was, for the most part, unsuccessful. Although the equation successfully explained performance when standardized test data were substituted for the criterion-referenced predictor, application to new data sets failed to predict performance as well as did the mean of the observations itself. It appears that new equations are needed for each data set, and at the present time the contribution of one set of predictive equations may be simply to guide the choice of terms that might be employed in another study. Even so, the payoff from this kind of knowledge when evaluative research is in an exploratory mode may be high. Knowing what variables to measure in accounting for school achievement is one step toward the goal set by Bruner (1964) for a theory of instruction that would guide evaluation.

GPO 809-783-10

Curriculum Comparison Study

Earlier it was observed that curriculum comparison studies are often criticized because they fail to explain why one curriculum produces more positive results than another. It was proposed that, were the curricula to be compared represented in the evaluative design by a parsimonious set of descriptors, one might learn what aspect of a program contributed to any observed difference. The usefulness of the procedures utilized in the present study, however, depended crucially upon there being a somewhat balanced arrangement of the curricular characteristics. In the present study, the data set used fell short of what was required in this respect. Variables did not span the X space in the desired fashion, and the observations on them were nested within curricula. The distribution of points in the X space, clustered as they were suggested that a straightforward analysis of variance would have been more appropriately utilized.

The reasons for the problem encountered are several: (1) the implementing school had a departmental policy which ensured identical implementation measures to be taken on all classes using the same curriculum in two of the three cases, and this fact was not understood until the measurements were calculated; (2) curriculum content factors will not vary within curriculum when the measurements are based upon the textbooks used, but will if taken on the materials presented in a given class by a given teacher. It should be noted, however, that clustering itself does not rule out the model building approach, as Daniel and Wood (1971) illustrate. A number of clusters larger than the three in the present study are required for successful application of the approach, however. Only very large field tests, such as the U. S. Office of Education's Cooperative Reading Studies (Bond & Dykstra, 1967), in which ten curricula in over 200 classes were compared, lend themselves to the approach proposed here.

Despite the inappropriateness of the data set for definitively concluding that a particular curricular variable is important in predicting performance, the findings warrant some observations. First, the statistical quality of the predictive equation was high in the case of equations predicting the mean and standard deviation. On the other hand, the skew equation behaved as unsatisfactorily as did those reported by Lohnes. Second, the distributional descriptors contributed significantly to the predictive equations. Mean criterion performance was a function of the mean and standard deviation of group baseline performance, while baseline skew predicted criterion skew. Finally, the curricular variables did make a difference, even though interpretation was difficult in the case at hand. It was apparent from the correlation matrix, and from some of the predictive equations, that an instructional step of greater scope or the longer curriculum decision unit negatively affected the mean and standard deviation of group performance. A curricular decision maker could conclude that, for eighth grade mathematics, smaller organization and content chunks might change the location and spread of the distribution of group performance.

APPRAISAL OF THE FINDINGS IN LIGHT OF THE GOALS OF THE STUDY

The Utility for Evaluation Purposes of the Variables in the Study

Two kinds of variables were of special interest in the present study. First, the contribution to the predictive equations of several variables describing instructional processes and content characteristics was investigated. How the descriptors of the shape of the distribution of group performance functioned in the regression equations was additionally of interest. Both types of variables were found to be useful predictors of achievement.

Of the four curricular variables three functioned satisfactorily in this initial screening. Rate adaptiveness, a measure of degree of individualization, contributed significantly to the equations in the curriculum tuning study. Curriculum decision unit size and structural granularity also appeared to make a difference, though the interpretation was confounded with the particular curriculum being studied. Puissance, as a measure of content difficulty, failed to discriminate among the various programs. The raters of puissance, all mathematics educators, reported great difficulty in classifying content according to Walbesser's matrix. Such problems have been encountered earlier in attempts to classify subject matter according to other schema, notably Bloom's (1956) taxonomy of the cognitive domain, and are manifest in the low interjudge reliabilities usually reported for such ratings.

In this exploratory study of the utility of curricular variables, their technical characteristics were not investigated. If the use of these variables is to become more widespread, their reliabilities need to be studied and means developed of sampling observations which are both practical and technically sound. For instance, the arbitrary manner in which content samples were drawn to measure puissance might be replaced with sequential sampling procedures which terminate when the coefficient stabilizes.

The distributional descriptor variables functioned in the manner that might have been expected from Lohnes' (1972a, 1972b) studies. Statistics other than the baseline mean did improve the predictive quality of the equation for the criterion mean, and taken together the set of equations developed for a given evaluation can be interpreted in terms of distributional changes. Thoughtful analysis of what distributional characteristics are desirable is needed to stimulate their wider use, however. Educators might, for instance, set goals for the performance of the highest and lowest 25% of the pupil population, and these goals could have implications for changes in the standard deviation and skew of the total group's performance.

Data Analysis as a Statistical Technique Useful in Evaluation

The application of data analytic procedures within a regression framework made a difference in the outcome of the reanalyses of the field test data. This fact is borne out through comparison of the fitted equation with initial and intermediate equations. Only one of the six equations fitted was not improved after residuals were inspected, outliers dealt with, and new terms added in a subsequent analysis. Evidence of the utility of the adjunctive analytic techniques is provided by the fact that the percent of variation accounted for rose as much as 13% from the initial to the final analysis. Against this positive evidence must be weighed the author's long search through the various data sets for a set of variables that were illustrative of a variety of the data analysis techniques. For instance, in the curriculum tuning study it is somewhat misleading to suggest that a number of data analytic techniques will usually contribute to the final results. In unreported examples, the techniques were often unrevealing, and the initial equation stood or was only slightly modified as a consequence of the use of residual analyses. If data analysis is to become a standard statistical technique in the evaluator's or other educational researcher's repertoire, more examples which best illustrate its use need to be found, and the user sensitized to the amount of unproductive effort that will be associated with much of his data analysis. Still, the cost in time of unproductive data analysis could be justified in terms of the contribution the procedures will make in some research cases, and in terms of the greater cost of seeking further the hypothesized results through expensive empirical study without first probing the initial data.

Comments about the particular data analytic procedures are warranted. First, the opportunity presented through data analysis to build the model by adding variables not included in the original design proved advantageous. School organization and teacher variables entered particular regressions after they were identified through inspection of the pattern of residuals. Post hoc addition of variables to the equation, however, points to the need for iterative experimentation to confirm the predictive contribution of such variables. Were data analyses employed earlier in a field testing sequence, or were follow through studies planned, the necessary iteration could occur in the field test sequence.

Second, s_q, the index estimating pure error from near neighbors, appeared to function irregularly with the data in hand. About 40% of the possible pairwise combinations of observations contributed to the estimation of the error term and many of these pairs of observations were, in fact, quite distant from each other in the X space. When the sample is small, the quantity 4N-10 is large in proportion to $\binom{N}{2}$, and the estimate is less likely to differ greatly from the mean square residual than it might were the sample large.

Finally, the analysis of residuals appears to have promise for quality control of educational data. While the data reanalyzed in the present study were too old for the cause of specific aberrations to be pinpointed, old suspicions were aroused through the queries made about large residuals, and one was left with the impression that immediate consideration of extreme values could have uncovered problems which would have justified new measurements being taken.

A Prototype for Evaluative Research on Educational Achievement

The basic thrust of an evaluative research model which relates educational outcomes to pupil characteristics and educational processes has been described by Astin and Panos (1971) and Cooley (1971). The results of the present study confirm the usefulness of this approach in providing evaluative information of use to educators and curriculum designers. In particular, when the characteristics of groups and of the curricula are related through regression analysis to statistical descriptors of group performance, evaluators may determine how curricular treatments affect the achievement distribution. This specific prototype for evaluative research has been shown to yield information useful in assessing the effects of individualized instructional practices on achievement.

Frequently evaluations have yielded inconclusive results no matter what evaluative model was followed. Dissatisfaction with such outcomes has grown as studies without positive findings have accumulated while large sums were being expended for curriculum development activity and the evaluation itself. The paucity of evaluative findings is a problem, however, to which data analysis offers a partial solution, for with additional statistical tools the evaluator may probe deeper for meaning in his data. The joint use of data analysis and a statistical model that includes a more complete description of baseline and criterion achievement as well as quantifiable measurements on a variety of proven predictor variables should not only improve the record of conclusive evaluative findings but additionally lead to findings of import beyond the particular evaluation with which it is associated. Timely and iterative use of the methodology illustrated in the present study may further demonstrate its utility not only in increasing knowledge about how instruction achieves its effect but also in improving the achievement of students.

References

Anscombe, F. J., & Tukey, J. W. The examination and analysis of residuals. Technometrics, 1963, 5, 141-160.

Astin, A. W., & Panos, R. J. The educational and vocational development of college students. Washington, D. C.: American Council on Education, 1969.

Astin, A. W., & Panos, R. J. The evaluation of educational programs. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D. C.: American Council on Education, 1971.

Atkinson, R. C. Ingredients for a theory of instruction. American Psychologist, 1972, 27, 921-931.

Berliner, D. C., & Cahen, L. S. Trait-treatment interaction and learning. In F. N. Kerlinger (Ed.), Review of research in education. Itasca, Ill.: F. E. Peacock, 1973.

Besel, R. A linear model for the allocation of instructional resources. Paper presented at the meeting of the Institute of Management Sciences, Detroit, September-October 1971.

Bloom, B. S. (Ed.) Taxonomy of educational objectives: Handbook I, cognitive domain. New York: McKay, 1956.

Bond, G. L., & Dykstra, R. The cooperative research program in first-grade reading. Reading Research Quarterly, 1967, 2, 5-141.

Box, G. E. P., & Draper, N. R. A basis for the selection of a response surface design. Journal of the American Statistical Association, 1959, 54, 622-654.

Bracht, G. H. Experimental factors related to aptitude-treatment interactions. Review of Educational Research, 1970, 40, 627-645.

Bracht, G. H., & Glass, G. V. The external validity of experiments. American Educational Research Journal, 1968, 5, 437-474.

Brogden, H. E., & Taylor, E. K. The theory and classification of criterion bias. Educational and Psychological Measurement, 1950, 10, 159-186.

Bruner, J. S. Some theorems on instruction stated with reference to mathematics. In E. R. Hilgard (Ed.), Theories of learning and instruction: The sixty-third yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1964.

Carroll, J. B. A model of school learning. Teachers College Record, 1963, 64, 723-733.

Clausen, R. E. Studies instrumental to the Wisconsin plan for evaluation of driver education. Paper presented at the Regional Driver Training Workshop, Madison, Wisconsin, September 1971.

Cook, W. W. The functions of measurement in the facilitation of learning. In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951.

Cooley, W. W. Methods of evaluating school innovations. Paper presented at the meeting of the American Psychological Association, Washington, D.C., September 1971.

Cronbach, L. J. Course improvement through evaluation. Teachers College Record, 1963, 64, 672-683.

Cronbach, L. J., & Snow, R. E. Final report: Individual differences in learning ability as a function of instructional variables. Stanford, Calif.: Stanford University, 1969.

Daniel, C. Use of half-normal plots in interpreting factorial two level experiments. Technometrics, 1959, 1, 311-341.

Daniel, C., & Wood, F. S. Fitting equations to data: Computer analysis of multifactor data for scientists and engineers. New York: Wiley, 1971.

DeVault, M. V., Golladay, M. A., Fox, G. T., Jr., & Skuldt, K. Descriptor for the analysis of individualized instruction. Madison, Wisc.: Wisconsin Center for the Analysis of Individualized Instruction, 1973.

Dixon, W. J. Simplified estimation from censored normal samples. Annals of Mathematical Statistics, 1960, 31, 385-391.

Draper, N. R., & Smith, H. Applied regression analysis. New York: Wiley, 1966.

Fisher, R. A. Moments and product-moments of sampling distributions. Proceedings of the London Mathematical Society, 1928, 30, 199-238.

Flanagan, J. C., Mager, R. F., & Shanner, W. M. Science behavioral objectives: A guide to individualized learning. Palo Alto: Westinghouse Learning Press, 1971.

Gage, N. L., & Unruh, W. R. Theoretical formulations for research on teaching. Review of Educational Research, 1967, 37, 358-370.

Gagné, R. M. Domains of learning. Interchange, 1972, 3, 1-8.

Gorth, W. P. Comprehensive achievement monitoring: The National Center at UMass. The CAM Newsletter, 1972, 1972(6), 1.

Guilford, J. P. The nature of human intelligence. New York: McGraw-Hill, 1967.

Harris, M. L., & Stewart, D. M. Application of classical strategies to criterion-referenced test construction: An example. A paper presented at the meeting of the American Educational Research Association, New York, February 1971.

Kendall, M. G., & Stuart, A. The advanced theory of statistics. Vol. 1. Distribution theory. (2nd ed.) New York: Hafner, 1963.

Kendall, M. G., & Stuart, A. The advanced theory of statistics. Vol. 2. Inference and relationship. (2nd ed.) New York: Hafner, 1967.

Klausmeier, H. J., Quilling, M. R., Sorenson, J. S., Way, R. S., & Glasrud, G. R. Individually guided education and the multiunit elementary school: Guidelines for implementation. Madison, Wisc.: Wisconsin Research and Development Center for Cognitive Learning, 1971.

Klein, S. P. The uses and limitations of standardized tests in meeting the demands for accountability. Evaluation Comment, 1971, 2(4), 1-7.

Krathwohl, D. R. Stating objectives appropriately for program, for curriculum, and for instructional materials development. Journal of Teacher Education, 1965, 16, 83-92.

Lachman, R. The model in theory construction. Psychological Review, 1960, 67, 113-129.

Lindquist, E. F. Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1951.

Lohnes, P. R. Planning for evaluation of the LRDC instructional model. Pittsburgh: Learning Research and Development Center, 1972. (a)

Lohnes, P. R. Statistical descriptors of school classes. American Educational Research Journal, 1972, 9, 547-557. (b)

National Education Association of the United States of America. Resolutions and other actions. Atlantic City, N. J.: National Education Association Publications, 1972.

Otto, W., & Askov, E. The Wisconsin Design for Reading Skill Development: Rationale and guidelines. Minneapolis: National Computer Systems, 1972.

Page, E. B. Miracle in Milwaukee: Raising the IQ. *Educational Researcher*, 1972, *1*, 8-16.

Quilling, M. R., & Otto, W. *Evaluation of the Word Attack Element of the Word Attack Element of the Wisconsin Design for Reading Skill Development: A report on the large-scale field test*. (Technical Report) Madison: Wisconsin Research and Development Center for Cognitive Learning, in press.

Quilling, M. R., & Wojtal, P. *The Wisconsin Design for Reading Skill Development: Study Skills. A report on the pilot test: 1970-72.* (Working Paper No. 97) Madison: Wisconsin Research and Development Center for Cognitive Learning. 1972.

Sanders, N. M. *Classroom questions: What kinds?* New York: Harper and Row, 1966.

Scriven, M. The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*. AERA Monograph Series on Curriculum Evaluation, Vol. 1. Chicago: Rand McNally, 1967.

Scriven, M. Pros and cons about goal-free evaluation. *Evaluation Comment*, 1972, *3*(4), 1-4.

Shedd, M. R. *Issues in implementation I*. Proceedings of the Conference on Educational Accountability. Princeton, N. J.: Educational Testing Service, 1971. (ERIC #051 313)

Skager, R. W. Objective based evaluation: Macro-evaluation. *Evaluation Comment*, 1970, *2*(2), 7-10.

Skager, R. W. The system for objectives-based evaluation--reading. *Evaluation Comment*, 1971, *3*(1), 6-11.

Stake, R. E. The countenance of educational evaluation. *Teachers College Record*, 1967, *68*, 523-540.

Stake, R. E. An approach to the evaluation of instructional programs (program portrayal vs. analysis). Paper presented at the meeting of the American Educational Research Association, Chicago, April 1972.

Stake, R. E. To evaluate an arts program. Unpublished manuscript, University of Illinois, 1973.

Stake, R. E., & Gjerde, C. *An evaluation of TCITY: The Twin City Institute for Talented Youth*, 1971. Urbana, Ill.: CIRCE, University of Illinois, 1971.

STATJOB summary: Reference manual for the 1108. (3rd rev.) Madison, Wisc.: University of Wisconsin, Academic Computing Center, 1972.

Stuffelbeam, D. L. Evaluation as enlightenment for decision making. In W. H. Beatty (Ed.), Improving educational assessment and an inventory of measures of affective behavior. Washington, D. C.: Association for Supervision and Curriculum Development, 1969.

Suppes, P. Some theoretical models for mathematics learning. Journal of Research and Development in Education, 1967, 1, 5-22.

Taba, H. Teaching strategies and cognitive functioning in elementary school children. Cooperative Research Project No. 2404. San Francisco: San Francisco State College, 1966.

Tukey, J. W. The future of data analysis. Annals of Mathematical Statistics, 1962, 33, 1-67.

Tyler, R. W. General statement on evaluation. Journal of Educational Research, 1942, 35, 492-501.

Vaughn, K. W. Planning the objective test. In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951.

Walbesser, H. H. Behavioral objectives, a cause celebre. The Arithmetic Teacher, 1972, 19, 418, 436-440.

Wang, M. C., Resnick, L. B., & Schuetz, P. R. PEP in the Frick Elementary School: Interim evaluation report of the Primary Education Project, 1968-1969. Pittsburgh: Learning Research and Development Center, 1970.

Wetz, J. M. Criteria for judging adequacy of estimation by an approximating response function. (Doctoral dissertation, University of Wisconsin) Ann Arbor, Mich.: University Microfilms, 1964. No. 64-10331.

Wiley, D. E. Design and analysis of evaluation studies. In M. C. Wittrock & D. E. Wiley (Eds.), The evaluation of instruction. New York: Holt, Rinehart and Winston, 1970.

Wiley, D. E., & Bock, R. D. Quasi-experimentation in educational settings: Comment. The School Review, 1967, 75, 353-366.

GPO 809-783-11